

The international journal of science / 9 March 2023

outlook
Chronic kidney disease

nature

MURKY WATERS

Satellite images reveal rising tide of marine algal blooms

Artificial additives
Is bigger really better when it comes to AI language models?

Turning the tide
Six priorities to mitigate risk of flash floods on drylands

Drying up
Deforestation linked to reduced rainfall in tropical forests

Vol. 615, No. 7951
nature.com

VISIT…

LANZAROTE Caliente.com

The international journal of science / 9 March 2023

nature

Celebrate women in science – today, and every day

International Women's Day can serve to bring hope, highlight progress and inspire research communities to continue their efforts to push hard for true gender equality.

International Women's Day falls on 8 March; it aims to draw attention to women's achievements and the fight for gender equality. The day has its critics: too performative, say some; an opportunity for institutions to put on a facade of change by doing a photo shoot, say others, or to load over-burdened women in their organizations with yet more duties. But they are wrong. There is a need to raise awareness. Women in science still, on average, publish less and win fewer grants and promotions than do men. Harassment, assault and marginalization drive promising researchers out of science, especially those whose race, ethnicity, disability or sexual orientation makes them targets for discrimination.

Activism and action can engender change — if the systems that have long oppressed researchers who are not male can be made to shift. Part of achieving that goal involves raising awareness of what is possible if barriers are broken down. "We don't need another massively shiny campaign," says Jess Wade, a physicist at Imperial College London and a campaigner for gender equality. "We actually need to support the women scientists that we have."

It's in that spirit that *Nature* asked six women researchers how they will be celebrating International Women's Day. Martina Anto-Ocrah is an epidemiologist at the University of Pittsburgh in Pennsylvania who researches sex and gender disparities with an emphasis on women's health and global health, particularly in sub-Saharan Africa. Anto-Ocrah says she wants to celebrate the contribution of social scientists to the advancement of gender equality. Social scientists "are the people who highlight all the cultural issues in our society that hold women back", she says. One example is how, during the COVID-19 pandemic, publication rates for women scientists dropped more markedly than did those for men — confirming that women shouldered a greater share of responsibilities during that time, such as caring for families, leaving less time for research (E. B. Madsen *et al. eLife* **11**, e76559; 2022).

Sandra Díaz is an ecologist at the National University of Córdoba in Argentina and one of the leaders of IPBES, the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. Díaz wants to celebrate women as science leaders. Although they make up far from half of the researchers running major laboratories or winning big awards, women are increasingly realizing that they can be at the cutting edge of discovery and knowledge production, Díaz notes. "More and more women take leading roles in coming up with groundbreaking ideas, spearheading really risky scientific endeavours, or leading large science-policy bodies." She says girls are learning that, for a woman, engaging in a scientific career does not necessarily mean working in the shadows as a follower.

> **"It takes courage for women to enter into and persist in the scientific workforce."**

Wade will be lauding the success of programmes designed to support early-career researchers, such as the Rising Stars scheme at the Massachusetts Institute of Technology (MIT) in Cambridge, and the many similar schemes it has inspired around the world. Rising Stars offers mentoring and support for researchers from historically marginalized or under-represented groups, as they move through their careers. The programme is a direct result of the work of MIT biologist Nancy Hopkins, who showed the institute's male leadership that women had less lab space, lower salaries and fewer grants compared with men. This activism led the institute down a path of change, as a new book, *The Exceptions* by Kate Zernike, describes. Wade also writes and edits Wikipedia articles on scientists from historically marginalized groups. She finds that, when she creates or edits a page for a US-based scientist, very often they will have gone through a Rising Stars-style programme.

Tanya Monro, Australia's chief defence scientist, will be celebrating courage. "It takes courage for women to enter into and persist in the scientific workforce," says Monro, who works in a field that is more male-dominated than most. "It takes courage for women to speak out when they are expected to shoulder the brunt of changing scientific workplaces, so that the girls and women that follow them have better odds of thriving," she adds. "Women and girls in science need every ounce of that courage for themselves, to overcome the lingering confidence gap many of us carry through life. And I'm glad that these women persist, because it is that courage that allows them to contribute to creating the knowledge and impact that shapes our world."

Gihan Kamel, a physicist at SESAME, the Synchrotron-light for Experimental Science and Applications in the Middle East, based in Jordan, will also be celebrating breaking barriers. "There is progress," she says. "Not least in breaking the extremes in cultural and religious traditions and rules that are made by society and forced on women — usually inherited from one generation to another."

Aster Gebrekirstos, a senior scientist at the World Agroforestry Centre in Nairobi, will be marking women who have succeeded in their roles despite facing significant challenges. These challenges include wars and conflict, says Gebrekirstos, who is from Tigray, a region of Ethiopia that has been at the centre of a devastating conflict. The United Nations Economic Commission for Africa has published *Earth, Oceans and Skies* (www.africanwomenscientists.com), an open-access anthology of writing from and about African women scientists (including Gebrekirstos). The book is honest about the hardships women have endured "to reach where we are", Gebrekirstos says. Take a moment to acknowledge those hardships and to advance equity – today, tomorrow and every day after that.

© 2023 Springer Nature Limited. All rights reserved.

nature

'Elite university' strategies might boost rankings – but at what cost?

Japan plans to share US$2 billion annually between a handful of top universities. Experience elsewhere suggests the negatives of this approach might outweigh the positives.

Japan is an exemplar in global research funding. It is one of a small cluster of countries, including Israel, South Korea and Sweden, that consistently spend upwards of 3% of gross domestic product on research and development (R&D). The 2020 average for countries in the Organisation for Economic Co-operation and Development comes in at just under 2.7%.

But investment alone doesn't guarantee R&D success, and the nation's policymakers worry that not enough of Japan's top universities have the kind of autonomy afforded to their counterparts in Western nations, as Nature Index reports in a Japan-focused issue this week (see page S86). Officials are also concerned about the fact that Japan's universities have been dropping in international rankings.

The government's response, announced in 2021, is to create a ¥10-trillion (US$75-billion) national endowment fund for research. This will be invested, with the annual returns – expected to be around ¥300 billion – distributed between a select group of institutions. These will be called universities for international research excellence.

Japan's move to separate a subset of its universities is not unprecedented. Indeed, it comes at a time when prominent institutions elsewhere, such as those of the US Ivy League, Australia's Group of Eight and the United Kingdom's Russell Group, are having to think harder about the consequences of their exclusivity, both positive and negative – particularly when it comes to diversity, equity and inclusion.

The inspiration for Japan's fund comes, at least in part, from the endowment model of private institutions in the United States, but there are key differences that must be borne in mind. Harvard University in Cambridge, Massachusetts, for instance, has the world's largest endowment, totalling $50.9 billion at the end of the fiscal year 2022, and this delivers an annual payout to the university of around $2 billion. Some of this is used for research. But Harvard's endowment is made up of contributions from more than 14,000 individual donors, whereas Japan's fund is a public endowment. As Nature Index reports, this is fuelling concerns about possible political interference (see page S84), which risks undermining the ambition for more autonomy.

> **"Japan's fund is a public endowment, fuelling concerns about possible political interference."**

Eligible Japanese universities must apply by the end of this month. A board of expert advisers to the nation's Ministry of Education, Culture, Sports, Science and Technology will select the candidates by the autumn. But it is the government, not the research community, that will make the final call, which is far from ideal.

In the United Kingdom, where the practice of concentrating some research funding is well established, it has helped to boost the outputs and profiles of recipient institutions. Members of the Russell Group (comprising 24 research-intensive universities) receive around two-thirds of one of the main sources of public research funding, known as quality-related (QR) funding. This means that the country's roughly 130 remaining eligible universities share just one-third. In 2022, the QR pot exceeded £2 billion (US$2.4 billion).

Reactions to Japan's move have been mixed. Some researchers and policy analysts welcome the country taking a well-trodden path. Others argue that if redistribution of funding is needed, universities in Japan's poorer parts should benefit more, and that this would tap into previously unexplored avenues for boosting economic growth.

It's not often that countries shake up their funding arrangements so radically, but Germany offers an example. In 2005, the nation ushered in its Excellence Initiative, at least in part to distribute more funding among a subset of its largest universities. France has wanted to move in a similar direction, but researchers have resisted the idea.

Predictably, the German initiative has widened the funding gap between those institutions that have received excellence money and those that have not (L. Mergele and F. Winkelmayer *High. Educ. Policy* **35**, 789–807; 2022). And, after more than a decade, the initiative is changing. In its latest iteration, comparatively more funding is going to smaller universities and to interdisciplinary research. The United Kingdom is also facing calls for its next Research Excellence Framework – the scoring system for assessing research and distributing QR funding – to give greater weight to measures of research culture and the quality of a research environment.

Japan's chosen universities must increase the value of their research, develop new disciplines, target global issues and gain worldwide visibility. Loss of visibility is evident among Japan's top universities, some of which have dropped in the rankings in recent years. In 2013, 5 Japanese universities made the top 200 of the Times Higher Education's World University Rankings; a decade later, that figure has dropped to just 2. Concentrating research funding at a handful of institutions is one way to help universities achieve league-table positions, because some ranking methodologies give weight to funding for research, alongside staff numbers and publication data.

Japan is embarking on a path that is likely to enrich a small number of universities and might well have positive results in areas including research outputs and rankings. But amid mounting evidence of rising levels of burnout and staff turnover – and low job satisfaction – in academia worldwide, the country's policymakers should consider both what might be gained and what might be lost by pursuing a strategy of concentrating funding in this way.

© 2023 Springer Nature Limited. All rights reserved.

World view



NATASCHA MEHRABI

By Zia Mehrabi

How will crises affect food systems? Ask a 'digital twin'

From COVID-19 to the war in Ukraine, virtual models could inform global food policy before emergencies unfold.

In April 2020, I sat outside with a colleague on a deserted campus amid a COVID-19 lockdown. Shelves were emptying in supermarkets across Europe and North America. Panic buying underpinned that, but worker illness and restrictions on mobility and trade were looming: it was not clear how these would create supply-side disruptions. As food-systems scientists, we agreed we simply didn't have the tools to answer that question.

In recent years, global food security has had shocks from the COVID-19 pandemic, the Russian invasion of Ukraine, extreme weather events and more. My colleagues and I scramble to find convincing answers quickly when development outfits, aid organizations and government think tanks come calling. What would be the impact of COVID-19 mobility restrictions on harvests in sub-Saharan Africa? How would Germany's move away from Russian gas affect global fertilizer production and use? How would heatwaves change the ability of the poorest people to afford food?

As we talked back in 2020, it was clear to both me and my colleague what we needed: next-generation food-systems models, hooked up to real-world data, that consistently capture patterns of food production, transport, processing and consumption, allowing stress-testing and real-time informed responses to systemic shocks.

Such representations of physical reality – 'digital twins' – already exist in some sectors. They are used in aerospace engineering to prepare for and respond to critical system failures, and in manufacturing to maintain product quality. The European Union is funding an initiative to develop digital twins of Earth systems to tackle climate change and aid protection of nature. Digital twins of food systems will – at least at first – have fewer data to draw on, because the ability to sense food flows varies with location, food type and stage of processing. But with advances in data generation, computing power and artificial intelligence (AI), I am convinced the time for these models is now.

Historically, major food emergencies such as the global price crises of 1972–75 and 2007–08 have triggered leaps in food-system modelling. But efforts have been fragmented. Economists can create complex models of prices and trade for policy analysis, but with limited geospatial resolution. Agronomists have excellent high-resolution renderings of crop production and yield, but these typically stop at the farm gate. What's missing are the details of how food flows through supply chains – for example, how lentils from Canada make their way to a restaurant in India.

Yet most of what's needed exists. Satellites can monitor crop production and track how yields respond to weather and other factors. Location and traffic data for major roads, railways, waterways and ports are available in many countries. With GPS, we can tell the speed at which goods travel. We know where people live and can quantify their demand for various foods using purchasing-power data, and we have granular information on international trade volumes for major commodities. Logistics researchers have modelled how food flows from producers to consumers for countries including the United States – models just waiting to be draped over representations of real-world infrastructure.

> **"We will be able to model the effects of multiple simultaneous stressors – disease outbreaks, conflicts or energy shocks."**

Putting together these pieces will bring immense benefits, allowing near-real-time assessment of the impact of extreme weather, export restrictions or labour shortages. We will be able to model the effects of multiple simultaneous stressors – disease outbreaks, conflicts or energy shocks – and assess how long it would take to re-route flows to alternative ports, offset losses and buffer shocks.

Technical challenges will need attention. Many maps of processing facilities and the agricultural workforce are incomplete, but machine learning and AI offer opportunities to fill these gaps. Integrating public and private data while ensuring producer and consumer privacy is also difficult, but can be done with current cryptographic tools.

Implementing digital twins will be an vast undertaking, and must be done right. On average, poorer nations have less capacity to build and use food-system models: they lack funds, they have less-mature monitoring systems and data infrastructure, and more of their economic activity happens in unrecorded informal markets. Global efforts must counter these inequities and democratize access to models.

Promoting data and model sharing will be key. Digital twins built behind closed doors by industry – particularly large commodity traders and the consultants and insurers that serve them – might not best serve the public interest or the resilience of the food system. Hoarding of privileged information is a real concern given increasing corporate power. Terrorism and security risks are considerable, if in-depth understandings of food-system fragilities are made public or inadequately protected from hacking.

Only through partnership between the public and private sectors, strong governance and clear oversight will the risks be mitigated and the full benefits realized. United Nations bodies could guide this effort, bringing together the organizations equipped to maintain the core components and setting rules for data sharing and governance.

To make digital twins of food systems a success, funding agencies must step beyond their silos and pay for initiatives on the same scale as large physics projects.

The world must stop putting off food-systems data science until emergency strikes. With real-time insights, we could see key fragilities in food systems before it's too late.

Zia Mehrabi is an assistant professor in environmental studies at the University of Colorado Boulder.
e-mail: ziamehrabi@gmail.com

© 2023 Springer Nature Limited. All rights reserved.

Research highlights

SUNSCREEN RECIPE THAT KEEPS SEAS SAFE FOR CORAL

A new sunscreen could protect fragile corals in addition to our skin.

Chemical sunscreens on the market today use organic molecules that absorb ultraviolet radiation to protect skin from damage and reduce the risk of cancer. These molecules are small enough to penetrate the skin, and have the potential to disrupt the body's hormonal activity. Sunscreen that washes off swimmers' skin into the ocean can bleach coral – possibly killing it.

To develop safer sunscreens, Yuan Zeng at Tsinghua University in Beijing and colleagues made a variety of molecules containing ring-shaped structures that absorb UV light; similar structures are found in conventional sunscreens. The researchers linked these molecules with water-soluble molecules to make several polymers. They screened these to find the one that absorbed the most UV light.

In laboratory tests, the winning material protected mice from UV-induced skin burn better than commercial sunscreens. It did not irritate or get absorbed into the rodents' skin, and did not bleach corals.

The polymer is not biodegradable, however, and the authors say that future research could explore ways to make versions that are.

Cell Rep. Phys. Sci. https://doi.org/grvh57 (2023)



US COASTS FACING MORE FREQUENT DOUBLE CYCLONES

In coming decades, the US eastern and Gulf coasts face a growing risk of being hit by two hurricanes in quick succession – a development brought on by climate change.

A one-two punch of one hurricane followed quickly by another causes extra devastation. In 2021, for example, parts of Louisiana flooded when Hurricane Nicholas struck soon after Hurricane Ida unleashed heavy rains.

To understand how frequently pairs of hurricanes might strike in the future, Ning Lin at Princeton University in New Jersey and her colleagues analysed data on hurricanes that made landfall in the United States between 1949 and 2018. The scientists then simulated how storms could evolve by the end of the century.

For seven of nine coastal cities studied, the risk of two hurricanes striking within 15 days of each other has risen since 1949. In a scenario in which greenhouse-gas emissions rise at a moderate pace, back-to-back hurricanes might occur every one to 3 years by the end of the century, as opposed to every 10–92 years now.

Cutting greenhouse-gas emissions can cut the risk of these compounding hazards.

Nature Clim. Change https://doi.org/grt9bz (2023)

BONING UP: THE STEM CELLS THAT GIVE MINI ANTLERS TO MICE

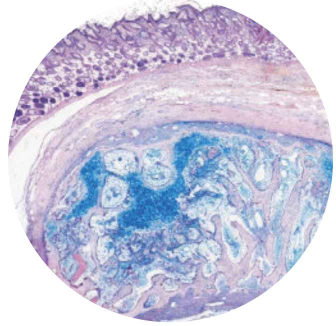
Every spring, deer lose their antlers; by early autumn, they have a new set. This bony headgear sprouts by about 2.75 centimetres a day, one of the fastest rates of bone growth among mammals.

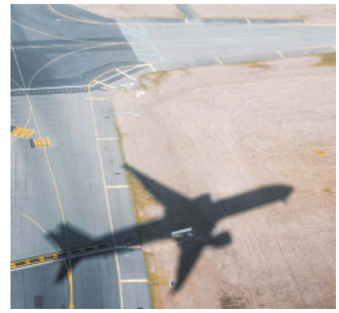
Tao Qin at Northwestern Polytechnical University in Xi'an, China, and his colleagues used RNA sequencing to examine almost 75,000 cells from sika deer (*Cervus nippon*), before, during and after the antlers were lost. The authors identified a specific family of stem cells that was key to antler renewal.

Ten days before the antlers were shed, one subtype of these stem cells was abundant in the stumps that remain on the day of shedding. By day five after shedding, these cells had generated a separate subtype of stem cell. By day ten after shedding, the new subtype of stem cells in the stumps began to develop into cartilage and bone cells. When the team transplanted day-five stem cells to the foreheads of mice, the rodents grew antler-like structures (pictured) within 45 days.

These stem cells could lead to treatments for bone or cartilage injuries, the authors suggest. However, this would raise safety, ethical and legal concerns.

Science **379**, 840–847 (2023)





FEMALE SCIENTISTS MISS OUT ON CAREER ADVANCES ABROAD

Researchers often move to another country to advance their careers, but this opportunity isn't afforded to men and women equally. Female researchers are less internationally mobile than their male counterparts, an analysis finds – but this gender gap has shrunk.

Xinyi Zhao at the Max Planck Institute for Demographic Research in Rostock, Germany, and her colleagues used data on 33 million publications from 10 million researchers to track researchers' international movements from 1998 to 2007. The team found that women were under-represented among those whose affiliations changed from one country to another, but this is shifting.

The number of internationally mobile female researchers almost tripled during the study period, from roughly 29,000 in 1998–2002 to 79,000 in 2013–17, compared with less than doubling for male researchers, from roughly 92,000 to 167,000. And the share of women among internationally mobile researchers increased faster than the share of women among all researchers.

Still, compared with men, women tended to travel shorter distances, and to travel to and from a narrower selection of countries.

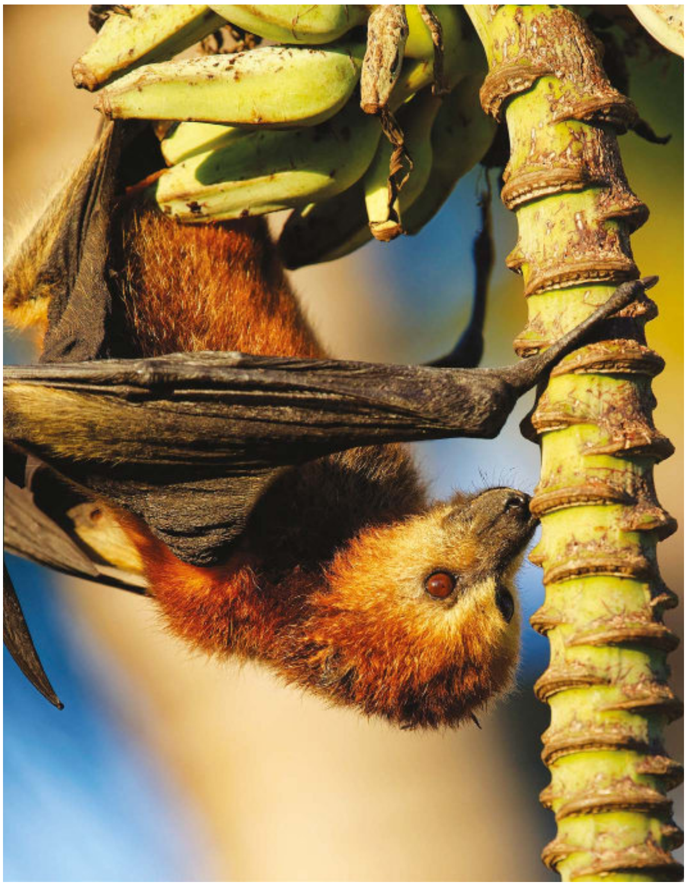
Proc. Natl Acad. Sci. USA **120**, e2214664120 (2023)

L TO R: EMILY KASK/AFP/GETTY; T. QIN *ET AL./SCIENCE*; FABIAN KRAUSE/EYEEM/GETTY

© 2023 Springer Nature Limited. All rights reserved.

For the latest research published by *Nature* visit: www.nature.com/latestresearch

HUNTING IS DRIVING LARGE BATS TO EXTINCTION



The Mauritian flying fox (*Pteroptus niger*) is a relatively large bat.

Large tropical bats with narrow home ranges are disproportionately likely to be hunted by humans, according to a global analysis of 1,320 bat species.

Krizler Tanalgo at the University of Southern Mindanao in Kabacan, the Philippines, and his colleagues found that 19% of the bat species they surveyed are hunted. Bats that live on nectar and other plant products are the top targets, especially the fruit-eating flying foxes and other members of the superfamily Pteropodoidea, along with some of the Rhinolopoidea, or horseshoe bats. Other key predictors that a species is likely to be hunted include a habitat that is easily accessible to humans, and low incomes among the hunting population.

The team found that hunted species are, in general, at higher risk of extinction than those that aren't hunted, suggesting that unsustainable hunting is a driver of population decline. Thus, the authors call for investments in creating economic security and reliable sources of meat from domesticated animals for people living in key bat habitats. Efforts to limit bat hunting could both protect bats and prevent the transmission of infectious diseases from bats to humans.

Biol. Conserv. **279**, 109944 (2023)

SHARPSHOOTERS ARE MASTERS OF SUPERPROPULSION

Sap-sucking insects called sharpshooters eject up to 300 times their body weight in liquid waste each day thanks to an energy-saving technique known as superpropulsion.

Sharpshooters (Cicadellidae) are millimetre-scale insects whose diet of plant sap is extremely poor in nutrients. To survive, they need to consume a huge amount of sap and, in turn, constantly and efficiently expel their liquid waste. Until now, the physics underpinning this excretion process had been unclear.

Elio Challita at the Georgia Institute of Technology in Atlanta and his colleagues show that the excretion occurs through superpropulsion – an effect in which an oscillating surface ejects an elastic projectile, such as a liquid droplet, at a speed higher than the greatest upward speed of the surface itself. In the case of sharpshooters, this surface is located on a pointy appendage called the anal stylus.

The team found that for these tiny insects, waste disposal by superpropulsion is much more energy-efficient than alternative mechanisms. The researchers say that their results could guide insect-inspired designs of self-cleaning structures and soft, elastic robotic engines.

Nature Commun. **14**, 860 (2023)



SECRETS IN THE DUST: SIGNS THE SUN HAD MANY SIBLINGS

The Sun was born within a cluster of stars. Seeking to determine the number of stars in that birth cluster, Sota Arakawa at the Japan Agency for Marine-Earth Science and Technology in Yokohama and Eiichiro Kokubo at the National Astronomical Observatory of Japan in Tokyo focused on aluminium-rich minerals found in some of the oldest dust particles in the Solar System.

According to one theory, the isotope aluminium-26 was injected into the Solar System by a supernova – an exploding star (pictured, artist's impression). The authors suggest that this injection must have occurred during the first 100,000 years of the Solar System's history.

The scientists used modelling to determine the probability of a supernova occurring in a given period of time. This allowed them to calculate how many stars must have been in the cluster to make it reasonably likely that one exploded at the appropriate time to supply the aluminium in the dust. The answer: 2,000–20,000 stars, depending on how long it took the Solar System to form.

Previous researchers estimated that the birth cluster held just 500 stars.

Astron. Astrophys. **670**, A105 (2023)

L TO R: FABRICE BETTEX PHOTOGRAPHY/ALAMY; PETER JURIK/ALAMY

© 2023 Springer Nature Limited. All rights reserved.

News in focus



SALWAN GEORGES/*THE WASHINGTON POST* VIA GETTY

A 7.8-magnitude earthquake struck Turkey and Syria on 6 February, killing more than 50,000 people.

ONE MRI FOR 4.7 MILLION PEOPLE: THE BATTLE TO TREAT SYRIA'S EARTHQUAKE SURVIVORS

英文杂志QQ群：855583774

With only 64 X-ray and 73 kidney-dialysis machines, 7 CT scanners and one MRI machine, doctors in northwest Syria are racing against the clock to treat 8,500 injuries.

By Miryam Naddaf

The 7.8-magnitude earthquake that struck Turkey and Syria on 6 February has killed more than 50,000 people and flattened multiple cities. More than 4,500 of those who died, and 8,500 of those injured were in northwest Syria, a region with no unified government that has been cut off from the world for more than 12 years amid a devastating war. Its already-depleted health-care system — 4.7 million people share just one MRI machine — is on its knees. More than 10,000 buildings are completely or partially destroyed, leaving 11,000 people homeless, according to the United Nations Office for the Coordination of Humanitarian Affairs. The earthquake also destroyed warehouses that store medicines. *Nature* spoke to doctors, engineers and other experts on the ground, as well as those helping remotely from Europe and from elsewhere in the Middle East. This is what they have said.

After the 2011 Arab Spring, Syria's government used military force against all opposition. Around 60% of the 4.7 million people in northwest Syria are internally displaced, having fled bombing attacks by the government of Bashar al-Assad, which has military backing from Russia. The UN and aid agencies are struggling to get supplies and expertise to earthquake-affected areas, because there are only three temporary crossing points along the 911-kilometre border with Turkey (see 'Crossing continents').

Medical emergency

Hospitals in this region have been overwhelmed as they attempt to help thousands of injured people in spaces with severely limited

© 2023 Springer Nature Limited. All rights reserved.

beds, medical supplies, surgical equipment and intensive-care facilities.

More than 8,500 injured people need to be accommodated in only 66 functional hospitals, which are providing 1,245 beds for short hospital stays, according to the World Health Organization (WHO). Moreover, all of northwest Syria has just 86 orthopaedic surgeons, 64 X-ray machines, 7 computerized tomography (CT) scanners and one magnetic resonance imaging (MRI) machine, according to WHO-compiled data published towards the end of 2022. Gynaecologist Ikram Haboush, director of the sole public maternity hospital in the city of Idleb, says "the medical situation in northwest Syria is catastrophic".

"The supply of antibiotics ran out from day three" after the earthquake, says Abdulkarim Ekzayez, an epidemiologist at King's College London, who now fears widespread infections. "We have used the medications and serums that would have lasted us for four to six months in two to three days," adds Haboush. The WHO has started airlifting medicines and medical supplies, but says northwest Syria also needs essential diagnostic equipment such as X-ray machines.

The region's critical lack of health care is partly caused by hospitals and medical staff being targeted during the war, explains Ekzayez, who is a co-investigator on a UK-funded project called Research for Health System Strengthening in Syria. In a separate study, Ekzayez estimated that bythis June 2021, 350 medical facilities had been attacked and 930 health-care personnel killed.

"People are running all over the place to make use of any existing resources, including basic ambulances," Ekzayez says. "The medical staff have been working non-stop," adds Haboush, who was working at the maternity hospital in Idleb when the first earthquake hit at 01:17 universal time. The hospital occupies the fifth and sixth floors of a building. "We had to evacuate the incubators and move all the babies to the ground floor," Haboush recalls.

According to three doctors whom *Nature* spoke to in northwest Syria, the most common injuries are limb fractures, trauma injuries, crush injuries including 'crush syndrome' (damages that result in organ dysfunction including kidney failure) and bleeding.

People with crush syndrome need intensive care and dialysis, says cardiologist Jawad Abu Hatab, who is dean of medicine at Free Aleppo University in northwest Syria. However, the region has only 73 renal-dialysis machines, according to the WHO-compiled data.

> **"The supply of antibiotics ran out from day-three after the earthquake."**

"There were also cases of cardiac arrests from the shock and horror of the catastrophe," Abu Hatab adds. The medical community is also preparing to deal with more cases of post-earthquake trauma, especially among children and women.

Doctors say that they urgently need more dialysis machines and orthopaedic-surgery equipment, along with painkillers and antibiotics. "These supplies were not abundant before the earthquake" and will now run out, says Haboush.

Naked-eye damage assessments

Volunteer engineers are going house-to-house checking which of some 8,500 earthquake-damaged buildings are safe enough for people to return to. But lacking in the necessary tools and safety equipment, they are resorting to tapping walls with household implements such as simple hammers and making decisions using the naked eye.

It is painstaking work. By 25 February, around 2,644 of the damaged buildings had been assessed, according to the Association of Free Syrian Engineers, a voluntary group based in A'zaz, near Aleppo, that is coordinating the effort.

At the same time, experts from the Syrian diaspora are helping with virtual assessments in places inaccessible to local engineers. Residents of damaged buildings are taking photos and videos of the interiors and sending them to members of the Syrian Engineers Association in Qatar, based in Doha.

"Assessing the structural status of buildings is as urgent as the medical situation and should not be underestimated because aftershocks are still ongoing," says Muhammad Azmie Tawackol, a civil engineer and the association's president. But in northwest Syria, people are having to use whatever methods are available, he adds.

The engineers are trying to estimate whether buildings are inhabitable (safe with minor cracks), temporarily unusable (needing reinforcement) or unsafe — in which case occupants must evacuate immediately.

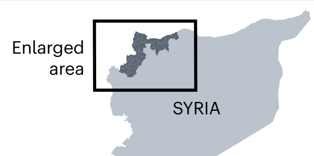
Buildings in danger of collapsing are being reinforced with whatever materials are available, irrespective of whether they are suitable. Carbon-fibre-reinforced polymers would work better for seismic reinforcement, Mohammad Khear Hayek, an engineer with the volunteers association, told *Nature*. Instead, they are having to use brittle industrial iron. "We are in an emergency situation, so we must respond quickly using the resources that we have," Hayek says.

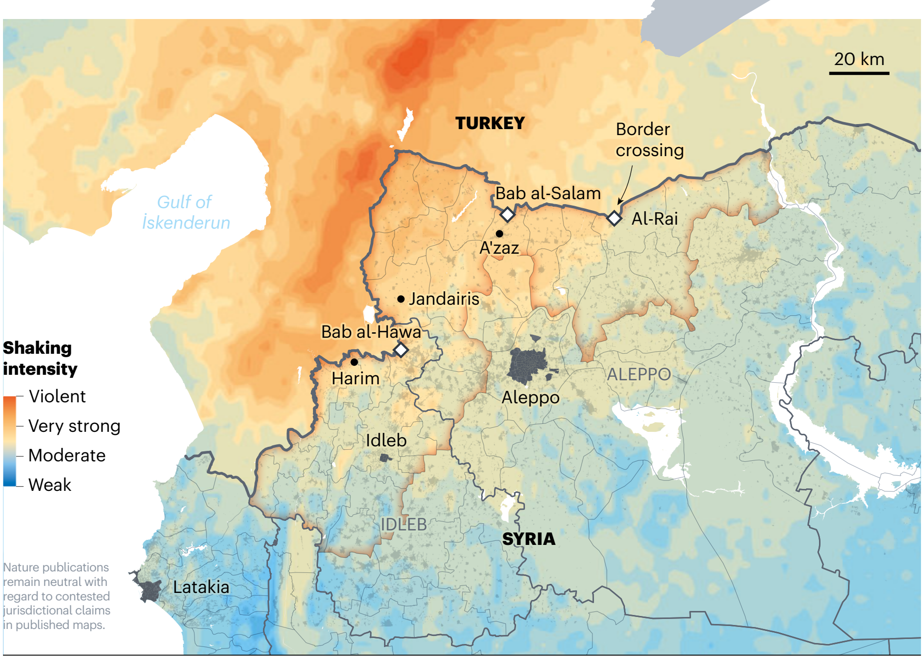
The engineers association based in northwest Syria is collating data daily and "the next step is to analyse the reports and produce statistical studies, which will be important in the reconstruction phase", says association member Ali Hallak, a computer engineer.

Many buildings will need to be rebuilt, but there is a shortage of engineers in northwest Syria, says Hayek. Before the earthquake, "we talked about a need to train engineers on reconstruction according to appropriate safety standards", Hayek says. "Now this has become a necessity," he adds.

CROSSING CONTINENTS

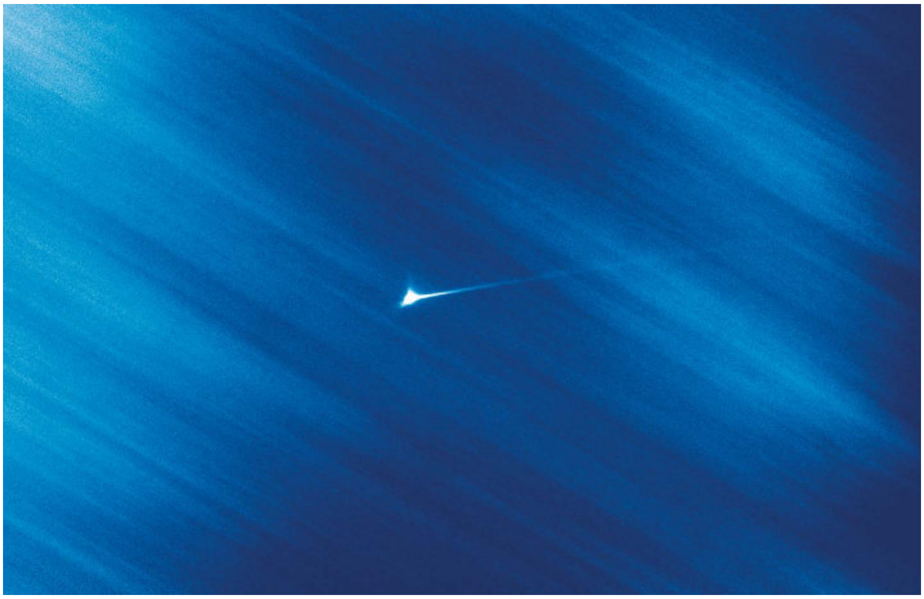
The 6 February earthquake killed more than 4,500 people in northwest Syria. Hospitals are overwhelmed from treating 8,300 injured people. Humanitarian aid is restricted to only three crossing points along the 900-kilometre border between Syria and Turkey.





SOURCE: EARTHQUAKE INTENSITY DATA, USGS; WATERWAYS, HUMANITARIAN OPENSTREETMAP TEAM; BUILDUP, ATLASAI; ADMINISTRATIVE BOUNDARIES, UN OFFICE FOR THE COORDINATION OF HUMANITARIAN AFFAIRS.

© 2023 Springer Nature Limited. All rights reserved.



UNIV. CANTERBURY ŌTEHĀWAI MT JOHN OBSERVATORY/UCNZ

The Didymos asteroid system, with stars appearing as streaks.

ASTEROID LOST 1 MILLION KILOGRAMS AFTER SPACECRAFT COLLISION

Studies reveal final moments before NASA's DART probe crashed into asteroid Dimorphos.

By Alexandra Witze

Last September, NASA's Double Asteroid Redirection Test (DART) spacecraft smashed into an asteroid, altering the rock's trajectory through space in a first test of planetary defence. Now scientists have deconstructed the collision and its aftermath — and learnt just how successful humanity's punch at the cosmos really was.

DART, which was the size of a golf cart, collided with a Great Pyramid-sized asteroid called Dimorphos. The crash caused the asteroid's orbit around another space rock to shrink — Dimorphos now completes an orbit 33 minutes faster than before the impact, researchers report[1] in *Nature*. If a dangerous asteroid were ever detected heading for Earth, a mission to smash into it would probably be able to divert it away from the planet.

DART's success has been reported before; now, five studies in *Nature* describe the final moments of the crash and how it affected the asteroid. One group combined data on the spacecraft's trajectory with photographs of the asteroid's surface just before impact[2]. As DART hurtled towards Dimorphos at more than 6 kilometres per second, the first part that hit was one of its solar panels, which smashed into a 6.5-metre-wide boulder. Microseconds later, the main body of the spacecraft collided with the rocky surface next to the boulder — and the US$330-million DART shattered to bits. The impact ejected at least one million kilograms of rock from Dimorphos's 4.3-billion-kilogram mass. The debris formed a tail that stretched for tens of thousands of kilometres behind the asteroid. Various telescopes watched over weeks as the tail shifted and evolved under the pressure of the Sun's rays; the Hubble Space Telescope even detected a second tail, which had disappeared by 18 days after the impact[3].

Dimorphos is 151 metres wide and orbits the larger asteroid Didymos. NASA's goal was to alter Dimorphos's orbit enough for astronomers to spot the changes by monitoring the brightness of the pair over time using ground-based telescopes. Neither asteroid is, or will ever be, a threat to Earth. Images taken as DART approached Dimorphos on 26 September show the asteroid covered in boulders. It seemed to be loose rubble barely held together by gravity — whose surface would probably shatter spectacularly when DART hit it.

Crash test

The discovery of more details is helping researchers to understand why the impact was so successful in shunting the asteroid, says Carolyn Ernst, a planetary scientist at the Johns Hopkins University Applied Physics Laboratory in Laurel, Maryland.

One factor is that the spacecraft hit a spot around 25 metres from the asteroid's centre, maximizing the force of its impact. Another is that large amounts of the asteroid's rubble flew outwards from the impact. The recoil from this force pushed the asteroid further off its previous trajectory. Researchers estimate that this spray of rubble meant Dimorphos gained almost four times as much momentum as was imparted by DART[4].

Although NASA has demonstrated this technique on only one asteroid, the results could be broadly applicable to future hazards, researchers say. "It means that we can quickly design a mission to deflect an asteroid if there is a threat, and we know that this has a very high chance of being effective," says Franck Marchis at the SETI Institute in Mountain View, California, who is also chief scientific officer at the telescope manufacturer Unistellar in Marseille, France.

"If you had asked me 30 years ago, 'Can we be confident we won't be wiped out by a giant killer asteroid a week from next Tuesday?' I would have had to say no," adds Tom Statler, DART's programme scientist at NASA headquarters in Washington DC. Now that astronomers have surveyed the skies to identify nearly all the dangerous asteroids — and now that DART has been shown to work —"we will know what to do about it when something new is found", he says.

Researchers are continuing to work through the DART data to learn more about the physics, chemistry and geology of both Dimorphos and Didymos. This work is being done with the help of a network of amateur astronomers co-led by Marchis. The network's members observed the asteroids with their telescopes before, during and after the impact. They discovered that the rocks seemed to become significantly redder immediately after the spacecraft hit[5].

A later colour change to blue was spotted by NASA's Infrared Telescope Facility in Hawaii, as reported by Cristina Thomas, a planetary scientist at Northern Arizona University in Flagstaff, at a meeting of the American Geophysical Union last December. "We think this is likely because we have a lot of material from Dimorphos thrown off," she says. The impact blasted through the asteroid's weathered interior and exposed part of its insides, making the asteroid appear temporarily blue, a few hours after impact.

1. Thomas, C. A. *et al. Nature* https://doi.org/10.1038/s41586-023-05805-2 (2023).
2. Daly, R. T. *et al. Nature* https://doi.org/10.1038/s41586-023-05810-5 (2023).
3. Li, J.-Y. *et al. Nature* https://doi.org/10.1038/s41586-023-05811-4 (2023).
4. Cheng, A. F. *et al. Nature* https://doi.org/10.1038/s41586-023-05878-z (2023).
5. Graykowski, A. *et al. Nature* https://doi.org/10.1038/s41586-023-05852-9 (2023).

© 2023 Springer Nature Limited. All rights reserved.

How to stop the bird flu outbreak becoming a pandemic

From tracking the disease's spread in wild birds to updating human vaccines, there are measures that could help keep avian influenza in check.


A goose being vaccinated against avian influenza in China.

WEI LIANG/CHINA NEWS SERVICE/GETTY

Fears are rising about bird flu's potential to spark a human pandemic, as well as its destruction of wildlife and farmed birds. An 11-year-old girl tragically died in Cambodia two weeks ago after catching avian influenza. That followed reports earlier this year of the virus spreading from mammal to mammal through a mink farm, and killing sea lions. Since the beginning of 2022, more than 50 million poultry birds in the United States, and a similar number in Europe, have either died of the disease or been killed in efforts to stem its spread. Can bird flu be stopped, and if yes, how?

Protecting poultry

Poultry farms are a key battleground in the fight against H5N1, the strain of avian influenza that is currently circulating. Outbreaks on farms threaten food security and provide opportunities for the virus to spread to farm workers. For decades, farmers have controlled the disease by culling infected animals. But now, with many countries experiencing outbreaks on dozens of farms every month, this is becoming untenable.

Some countries, including China, vaccinate poultry to limit the spread and severity of bird flu, and other governments around the world are now implementing vaccination policies or contemplating doing so. One problem with existing vaccines is that they cause birds to test positive for the virus, meaning farmers can't guarantee their birds are free of H5N1. This has "huge international trade and export implications", says Keith Poulsen, a specialist in infectious diseases who directs the Wisconsin Veterinary Diagnostic Laboratory in Madison.

Scientists are in the early stages of developing vaccines that might solve this problem. Microbiologist Adel Talaat at the University of Wisconsin–Madison and his colleagues are developing a vaccine that uses only a small part of the virus's DNA. Tests targeting other genetic regions could differentiate between birds that have been vaccinated and those that are infected.

Poultry farmers could also raise a wider variety of bird breeds to stop the virus, says Nichola Hill, an ecologist at the University of Massachusetts in Boston. In Asia, where farmers have a long history of navigating outbreaks of bird flu, some have switched to breeds that are less susceptible to the virus.

Conserving wildlife

H5N1 has become entrenched in wild bird populations over the past year, but there are "some small Band-Aids we can put on things", says epidemiologist David Stallknecht at the University of Georgia in Athens. Administering vaccines to wild birds is logistically difficult. So, for the most part, birds have to develop resistance to the disease through being infected, and many will die in that process.

Vaccines could help to protect certain species, Stallknecht says. Bald eagles (*Haliaeetus leucocephalus*), for example, can be severely affected by the virus, and some scientists are worried about the impact of bird flu on the population. But the strategy could only be used for species under grave threat, when "you're doing everything you possibly can to keep them on the planet", he says.

At the moment, Stallknecht and other wildlife researchers are trying to understand which species of wild bird are most severely affected by avian flu, and the implications this has for the spread of the disease. As well as helping scientists direct conservation measures, this research could give farmers a better idea of when bird flu might be heading their way if, for example, it is combined with when certain birds are known to migrate.

That knowledge could help farmers target measures to protect poultry, such as cleaning up grain that could attract wild birds and washing boots before entering farms. "It's extremely hard to do that 365 days of the year," Hill says.

Stopping a human pandemic

The death of the girl in Cambodia — and the fact that her father also tested positive for avian influenza — has renewed concerns about whether bird flu could spark widespread infection in people, or even a pandemic. "That's hard to say," says Thijs Kuiken, a veterinary pathologist at the Erasmus University Medical Center in Rotterdam, the Netherlands.

Ancestral versions of today's H5N1 virus have been circulating among birds for about 25 years and have not yet gained the ability to

© 2023 Springer Nature Limited. All rights reserved.



spread between humans. That leads Kuiken to think that the risk of a human pandemic is low. But the uptick in cases among wild birds and the discovery that the virus can be transmitted between mammals increase the risk that it could start spreading in humans. Kuiken would like to see increased surveillance of people who work in the poultry sector to make sure anyone who becomes infected is quickly isolated and treated.

If bird flu does trigger a human pandemic, there are a number of tools for combating the disease. Approved human vaccines against avian flu exist, and the World Health Organization monitors the evolution of H5N1 so that these vaccines can be updated appropriately. In the United States, the Biomedical Advanced Research and Development Authority has a stockpile of vaccines, although the supply is too low to be used to vaccinate widely around the world. Animal studies and observational data in humans suggest the antiviral drug Tamiflu is effective against H5N1 in people (J. R. Smith *J. Antimicrob. Chemother*. **65**, ii25–ii33; 2010), although there have been reports of resistant strains (M. D. de Jong *et al. N. Engl. J. Med*. **353**, 2667–2672; 2005). Non-pharmaceutical tools including face masks can also limit disease spread.

For a world still reeling from COVID-19, the prospect of another pandemic is alarming. Bird flu's current mortality rate in humans is around 50%, although that would be likely to drop if the virus gained the ability to infect cells in the upper respiratory tract — a prerequisite for efficient human-to-human spread. But several scientists say that an H5N1 pandemic would probably be more manageable than COVID-19 because of the drugs and vaccines that are already available, and because of tools such as mRNA vaccines that were developed as a result of COVID-19. "Not to say it won't be a mess," says Stallknecht, "but it probably won't be as bad as it could be."

Hill agrees that humanity has the tools needed to keep the virus in check. "The question is control at this point, and preventing a human pandemic," she says. "And I think both of those are achievable goals."

By Saima May Sidik

News explainer

ANCIENT GENOMES SHOW HOW HUMANS ESCAPED EUROPE'S DEEP FREEZE

A pair of studies offers the most detailed look yet at groups of hunter-gatherers during the last ice age.

By Ewen Callaway

Like retirees who flock to the Costa del Sol, ancient European hunter-gatherers sought out Spain's warmer climate during the peak of the last ice age.

A pair of ancient-genome studies shows that humans who had holed up in the Iberian Peninsula repopulated western Europe after the retreat of glaciers that covered large parts of the continent from about 26,000 to 19,000 years ago[1,2].

The research – based on newly sequenced genomes from more than 100 individuals – offers the most detailed look yet at groups of hunter-gatherers who lived in Europe before, during and after the last ice age.

Ancient genomes from this period are scarce, so "even adding one data point is really important", says Mateja Hajdinjak, a molecular biologist at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany, who was not involved in either study.

Homo sapiens migrating out of Africa reached Europe at least 45,000 years ago – and possibly earlier. But the handful of ancient genomes from this period suggest that these pioneers left no genetic trace in later hunter-gatherers.

Instead, a landmark 2016 study[3] identified a genetic signature in 35,000-year-old remains from the Goyet cave system in Belgium, which persisted in hunter-gatherer populations that lived tens of thousands of years later. The Goyet remains were associated with artefacts from the Aurignacian, a Europe-wide material culture known for its elaborate cave-wall art and 'Venus' figurines.

Missing persons

But the 2016 study raised a mystery. The Goyet ancestry was missing in remains from a roughly 20,000-year period leading up to and during the peak of the ice age, before reappearing later in hunter-gatherers in western Europe. "Where were these people hiding for 20,000 years?" asks Cosimo Posth, a palaeogeneticist at the University of Tübingen in Germany.

To find out, Posth and his colleagues sequenced ancient DNA from 116 hunter-gatherers who lived in Europe and western Asia between 35,000 and 5,000 years ago, and analysed previously sequenced genome data from hundreds more.



JÜRGEN VOGEL/LVR-LANDESMUSEUM BONN

These 14,000-year-old skulls were found in western Germany. Their genetic ancestry suggests that human populations migrated in response to Europe's changing climate.

© 2023 Springer Nature Limited. All rights reserved.

Among the trove they detailed in *Nature* on 1 March[1] are several pre-ice-age individuals from sites in France and Spain. Their remains were found with artefacts attributed to another pan-European material culture known for its figurines, called the Gravettian. When Posth's team looked at these people's genomes, the researchers found a direct link to the earlier Goyet population that had seemingly vanished.

The researchers traced this ancestry to 21,000- and 23,000-year-old remains from sites in northern Spain and southwestern France, respectively. These sites are linked to another culture called the Solutrean.

A separate study reported on 1 March in *Nature Ecology and Evolution*[2] found that the 23,000-year-old body of a man found at a site in southern Spain – also linked to Solutrean artefacts – also showed evidence of Goyet ancestry.

The two studies suggest that the Iberian Peninsula was a refuge for hunter-gatherers as the climate cooled and glaciers ensconced northern Europe. The genetic signature – the same one found in Goyet 35,000 years ago – later pops up across western Europe and even into Poland after Europe's climate warmed.

Ice-age refuges

Iberia wasn't Europe's only ice-age holdout, says Posth. His team discovered a genetic signature in post-ice-age people in northern Italy that eventually reached Sicily. But these people were not related to humans who were in Italy before the peak of the ice age, nor to the western hunter-gatherer groups.

Instead, they descend from people who saw out the coldest periods in southeastern Europe, probably in the Balkans, says Posth. There is currently no ancient DNA from this period to confirm that hunch, he adds, leaving a "big empty spot on the map".

Colin Wren, an archaeologist at the University of Colorado Colorado Springs, says those findings confirm his own work, suggesting that ice-age Italy was less hospitable to humans than was Iberia[4]. Italy was once connected to Croatia by a now-submerged plain, so it makes sense that eastern hunter-gatherers moved in, he adds.

The studies paint a dynamic picture of European hunter-gatherer populations, which don't always line up with cultural shifts, says Natasha Reynolds, a palaeolithic archaeologist at the University of Bordeaux, France. They also show that, for parts of Europe at least, the coldest period of the ice age wasn't as inhospitable as it's often made out to be. "People weren't just huddling in caves waiting for the glaciers to retreat," she says.

1. Posth, C. *et al. Nature* **615**, 117–126 (2023).
2. Villalba-Mouco, V. *et al. Nature Ecol. Evol.* https://doi.org/10.1038/s41559-023-01987-0 (2023).
3. Fu, Q. *et al. Nature* **534**, 200–205 (2016).
4. Wren C. D. & Burke, A. *PLoS ONE* **14**, e0217996 (2019).



Activists in Los Angeles, California, protest against anti-Asian racism.

DAVID MCNEW/AFP/GETTY

SHADOW OF US CHINA INITIATIVE LOOMS LARGE FOR SCIENTISTS

Anti-Asian scrutiny has only intensified since the programme ended one year ago, researchers say.

By Natasha Gilbert

One year after the US government ended its controversial China Initiative, scientists of Chinese heritage say that they are still being targeted unfairly and fear for their safety.

The initiative – which was aimed at safeguarding US laboratories and businesses from espionage – created the perception of bias against researchers of Chinese descent, said assistant attorney-general Matthew Olsen when shutting it down in February 2022, although he denied that the programme had actually used racial profiling. While it was active, more than 150 people were criminally charged for actions such as failing to disclose funding or partnerships with institutions in China, according to an analysis by *MIT Technology Review*. Nearly 90% of them were of Chinese heritage. Many of the charges brought by the US Department of Justice (DoJ) after the initiative's launch in 2018 were eventually dropped or dismissed, and some prosecutions ended in acquittal.

The climate of fear and anxiety hasn't gone away – researchers are just being pressured in a new way, says Jenny Lee, a social scientist at the University of Arizona in Tucson who studies research collaborations and geopolitics. Since the initiative's official shutdown, the US government has adopted various anti-China policies. And although the DoJ is pursuing fewer criminal charges, it says that it will work increasingly with federal agencies to investigate researchers and issue civil and administrative penalties for noncompliance. Universities are also taking a more active role in assisting investigations and pursuing potential wrongdoing, sources tell *Nature*.

Unfortunately, the scrutiny "has only intensified", says Gang Chen, a mechanical engineer at the Massachusetts Institute of Technology in Cambridge, who was arrested in January 2021 under the China Initiative, only for the DoJ to drop the charges a year later. He and others who have had their lives upended by the initiative have been speaking out about the damage that it has done.

"The government has not done enough" to ease the situation, Chen adds. The DoJ did not

© 2023 Springer Nature Limited. All rights reserved.



respond to *Nature*'s request for comment.

One example of a university taking a more active role in the initiative's wake was reported by *The San Diego Union-Tribune* in December 2022. Xiang-Dong Fu, a molecular biologist at the University of California, San Diego (UCSD), was forced to leave his position after the university accused him of hiding ties to China. UCSD said he had violated its conflict-of-commitment policy by accepting travel reimbursements from Chinese institutions that he had visited, and had failed to disclose Chinese grants that bore his name. Fu denies any wrongdoing, according to the *Tribune*.

Universities reject the idea that they are unfairly targeting researchers of Chinese heritage. According to Toby Smith, vice-president for science policy and global affairs at the Association of American Universities in Washington DC, US institutions acknowledge the considerable research contributions from these scientists. Universities are working to ensure that all faculty members are disclosing information properly, he adds.

But he calls on US funding agencies to provide greater clarity for universities on what counts as an offence and what are appropriate and fair sanctions.

Scientists need support, says Gisela Kusakawa, executive director of the Asian American Scholar Forum, a non-profit organization based in New York City. Universities and agencies should provide training for scientists on how to complete disclosure forms, and they must allow scientists the opportunity to revise completed forms to ensure that they are correct, she says.

The illusion of improvement

In the past year, the US government has adopted several policies and positions that have perpetuated the narrative that scientists from China are potential spies, Lee says. In August 2022, the US Congress passed into law the CHIPS and Science Act, which earmarks an extra US$280 billion for research and innovation and includes measures designed to tighten research security. For example, it asks US institutions to report gifts of $50,000 or more from a foreign government, down from the previous minimum of $250,000.

In January, Congress also voted to form a bipartisan committee to assess the economic and competitive threats that China poses to the United States. The Association of American Universities has said that the creation of the committee signals an intent in Congress to monitor China's influence on the nation's scientific enterprise.

The US government has caught genuine Chinese spies stealing trade secrets and scientific and technological developments. But many say that the government's broad-brush scrutiny of researchers of Chinese descent is excessive, and could actually harm national security by demonizing scientific collaboration with China and driving out scientists who contribute to US scientific prowess. In a survey published in February, Lee says, she found a link between fears of racial profiling and a desire among scientists to return to China (J. J. Lee and X. Li *Rev. High. Educ.* https://doi.org/jzw7; 2023).

"I am afraid of doing any research. We always live in fear."

The end of the China Initiative gave the illusion that researchers of Chinese heritage would be targeted less, she says, but "the chilling effect" is "still very much at play".

Afraid of doing research

Researchers unjustly accused under the China Initiative and now rebuilding their lives and careers are emblematic of this situation.

Xiaoxing Xi, a physicist at Temple University in Philadelphia, Pennsylvania, was arrested at gunpoint in front of his family by the DoJ in 2015. Although this was before the initiative launched officially in 2018, scrutiny of researchers of Chinese heritage had begun years earlier. Xi was accused of passing information to scientists in China about restricted technology. The DoJ eventually dropped the charges.

Xi has been seeking damages for harm he experienced as a result of his arrest, and is appealing against a ruling in March 2022 that dismissed his claims. He is nervous about applying for federal research funding, and spends much of his time following the cases of targeted scientists and giving talks to raise awareness about anti-Asian sentiment. Before his arrest, he was juggling 9 research projects and had 15 people working in his lab. Now, he is working on just one project and has one researcher on his team.

"I am afraid of doing any research," he says. "We always live in fear."

Chen is similarly afraid to apply for federal research funding, concerned that the government could misuse the forms against him as they did before, he says. To feel more secure, he has switched from researching nanotechnologies with obvious commercial applications to doing more-fundamental science, exploring the solar evaporation of water. He also rarely answers e-mails from researchers or students in China who write asking questions about his research papers.

Anming Hu, a nanotechnology researcher at the University of Tennessee in Knoxville, who was indicted for hiding ties with China in 2020 and put under house arrest for more than a year before being acquitted, is also trying to get his research back on track. He has spent the past year rebuilding his lab, but has had trouble securing any funding. Currently, he has two graduate students on his team; before his arrest he had six, he says, adding that he won't now take on students or researchers from China because it's too risky.

"If nothing had happened to me, I would have climbed to a much higher research level," he says. "But I do my best."

DATA HINT AT RUSSIA'S SHIFTING SCIENCE COLLABORATIONS

Nature analysis suggests that Russia is increasing partnerships with China and India.

By Richard Van Noorden

A year into Russia's war on Ukraine, its effects on global research collaborations might be starting to show up in the scientific literature.

Nature analysed co-authorship patterns on papers in the Scopus database. The results suggest that, last year, an increased share of Russia's internationally collaborative papers had co-authors from China and India, whereas the proportion co-written with US or German authors fell. Ukraine, meanwhile, has sharply reduced its scholarly ties with Russia, and seems to have increased research connections with Poland.

The data available so far can only hint at possible changes, which will become more apparent later this year. Many of the papers published in 2022 were submitted to journals well before Western institutions halted scientific partnerships with Russia in response to the full-scale invasion, which began on 24 February that year. And databases of scientific papers will continue filling up their 2022 collections for another month or two, owing to

© 2023 Springer Nature Limited. All rights reserved.

routine indexing delays, so the raw numbers of papers will continue to increase. (Relative proportions of co-authorship are unlikely to vary at this stage.)

Co-authorship changes

Last year, just over one-quarter of Russia's papers and review articles — as recorded in Scopus — were internationally co-authored, a similar proportion to the year before. China is now close to overtaking the United States and Germany to become Russia's leading research partner, and India has soared up the rankings to seventh place (see 'Russia's top partners'). Among Russia's top 25 collaborators, the only other countries that have increased their share of its international co-authorships in 2022 are Kazakhstan and Iran.

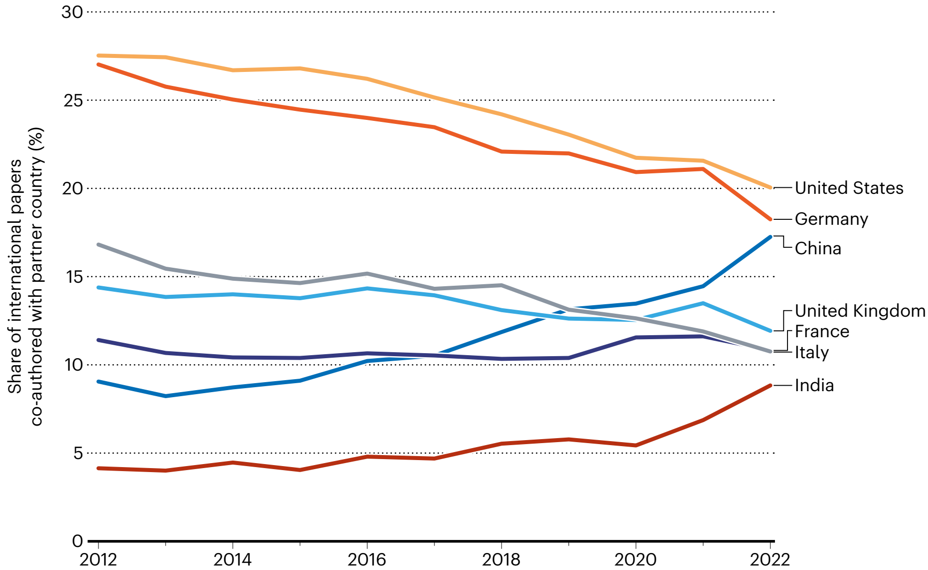
Kieron Flanagan, a science-policy researcher at the University of Manchester, UK, cautions that co-authorships are an imperfect proxy for real research-collaboration patterns, and that it's too early to draw conclusions from the available data. But he says that it does look as if Russia has seen a long-term relative decline in research collaborations with Western countries, and this trend could have accelerated in 2022.

China is increasing its scientific co-authorship share with most countries except the United States, so its rising influence on Russia might not be unusual, says Caroline Wagner, a science and policy researcher at the Ohio State University in Columbus. "The Russia–Ukraine conflict does not appear to have interrupted Russia and China's cooperation," she says.

In Ukraine, meanwhile, the proportion of papers with Russian co-authors fell sharply in 2022, and Poland is the country's clear leading research partner (see 'Ukraine's top partners'). Around 38% of Ukraine's papers in 2022 were

RUSSIA'S TOP PARTNERS

Last year, the share of Russia's international papers co-authored with researchers in China and India increased. Collaborations with the United States and Germany continued to fall.



NATURE ANALYSIS OF SCOPUS DATA

internationally co-authored — a slightly higher proportion than in previous years.

Nature found similar patterns in a parallel analysis of scientific papers in a different scholarly database, Dimensions (which differs from Scopus in its choice of journals to index). But according to the Dimensions data, Ukraine's collaboration with Poland has not increased: it has remained at around the same level since 2021.

Cutting research ties

Many research organizations cut collaboration activities with Russia shortly after its invasion of Ukraine. Some funders, including those in Poland and Germany, strongly discouraged researchers from continuing to co-author papers with Russian scientists. The German Research Foundation (DFG) — the country's main funding agency — said that this didn't apply to work that had been completed or submitted to journals before the invasion, and a DFG spokesperson adds that the agency is not monitoring its recommendations or introducing punishments for researchers who go against them.

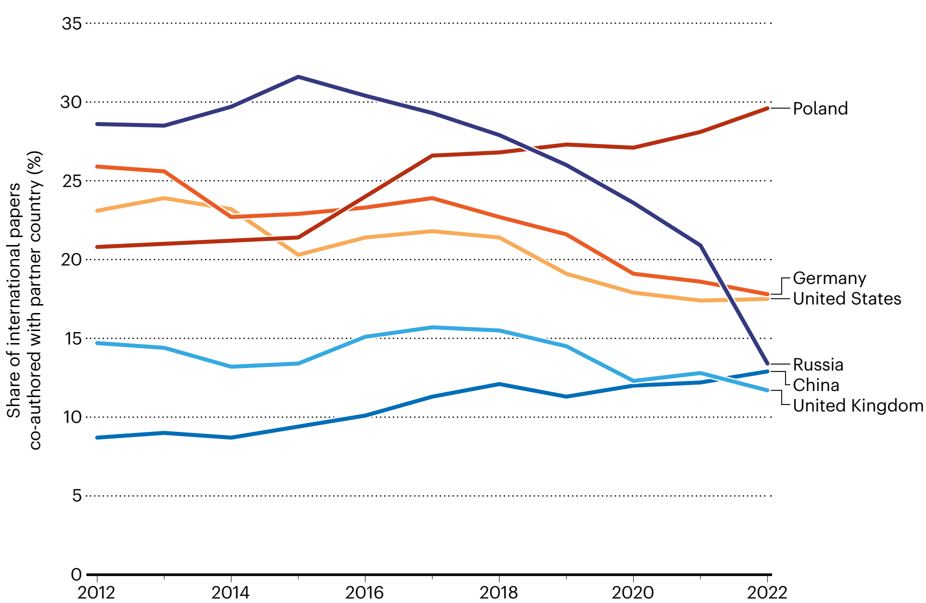
The research community has also discussed individual boycotts of Russian work — including in journals. International journals have generally not banned the consideration of Russian-authored work, and some researchers and journals have warned against indiscriminately isolating the country's scientists. But at least one title, Elsevier's *Journal of Molecular Science*, has said it will no longer consider manuscripts from scientists at Russian institutions. Its editor-in-chief, Rui Fausto, a chemist at the University of Coimbra in Portugal, says that the policy still stands, although *Nature*'s analysis found at least nine articles published in the journal that had Russian authors and had been submitted after the war began. (Fausto says that he is looking into them.)

The Ukrainian government has strongly discouraged collaboration with Russian researchers and publication in Russian journals, says Michael Rose, who studies the economics of science and innovation at the Max Planck Institute for Innovation and Competition in Munich, Germany.

The war has come during a period in which nations are becoming more aware of the competitive geopolitical aspects of research collaboration, Flanagan adds, with countries expanding export controls and introducing restrictions on overseas collaboration in activities that have been deemed sensitive to national security.

UKRAINE'S TOP PARTNERS

The proportion of Ukraine's international papers co-authored with Russia has fallen sharply.



NATURE ANALYSIS OF SCOPUS DATA

© 2023 Springer Nature Limited. All rights reserved.



ILLUSTRATION BY FABIO BUONOCORE

IN AI, IS BIGGER BETTER?

As generative artificial-intelligence models get larger, some scientists advocate for leaner, more energy-efficient systems. **By Anil Ananthaswamy**

Artificial-intelligence systems that can churn out fluent text, such as OpenAI's ChatGPT, are the newest darlings of the technology industry. But when faced with mathematical queries that require reasoning to answer, these large language models (LLMs) often stumble. Take, for instance, this algebra problem:

> A line parallel to $y = 4x + 6$ passes through (5, 10).
> What is the y-coordinate of the point where this line crosses the y-axis?

Although LLMs can sometimes answer these types of question correctly, they more often get them wrong. In one early test of its reasoning abilities, ChatGPT scored just 26% when faced with a sample of questions from the 'MATH' data set of secondary-school-level mathematical problems[1].

This is to be expected: given input text, an LLM simply generates new text in accordance with statistical regularities in the words, symbols and sentences that make up the model's training data. It would be startling if just learning language patterns could allow LLMs to mimic mathematical reasoning reliably.

But back in June 2022, an LLM called Minerva, created by Google, had already defied these expectations — to some extent. Minerva scored 50% on questions in the MATH data set[2], a result that shocked some researchers in artificial intelligence (AI).

"The chatter in the community is that this is really kind of astounding," says Sébastien Bubeck, a machine-learning specialist at Microsoft Research in Redmond, Washington.

Minerva had the advantage that it was trained on mathematics-related texts. But Google's study suggested another important reason the model did so well — its sheer size. It was around three times the size of ChatGPT.

The Minerva results hint at something that some researchers have long suspected: that training larger LLMs, and feeding them more data, could give them the ability, through pattern-recognition alone, to solve tasks that are supposed to require reasoning. If so, some AI researchers say that this 'bigger is better'

© 2023 Springer Nature Limited. All rights reserved.

strategy might provide a path to powerful AI.

But there are reasons to doubt this thesis. LLMs still make glaring errors, and some scientists suggest that larger models are merely getting better at answering queries that fall within the scope of the correlations in their training data, rather than acquiring an ability to answer entirely new questions.

The debate is now playing out on the frontiers of AI. Commercial firms have seen better results with bigger AI models, so they are rolling out ever-larger LLMs – each costing millions of dollars to train and run (see 'The drive to bigger AI models'). But these models have major downsides. Besides concerns that their output cannot be trusted, and that they might exacerbate the spread of misinformation, they are expensive and suck up huge amounts of energy.

Critics argue that, ultimately, big LLMs will never be able to mimic or acquire skills that allow them to answer reasoning problems consistently. Instead, some scientists say, smaller, more energy-efficient AI is the way to make progress – inspired, in part, by the way the brain seems to learn and make connections.

Big, bigger, better?

LLMs such as ChatGPT and Minerva are giant networks of computing units (also called artificial neurons), arranged in layers. An LLM's size is measured in how many parameters it has – the adjustable values that describe the strength of the connections between neurons. Training such a network involves asking it to predict masked portions of a known sentence and tweaking these parameters so that the algorithm does a little better next time.

Do this repeatedly over billions of human-written sentences, and the neural network learns internal representations that model how humans write language. At this stage, an LLM is said to be pre-trained: its parameters capture the statistical structure of written language that it saw during training, including all the facts, biases and errors in the texts. It can then be 'fine-tuned' on specialized data.

To make Minerva, for instance, researchers started with Google's Pathways Language Model (PaLM), which has 540 billion parameters and was pre-trained on a data set of 780 billion tokens[3]. A token can be a word, digit or some unit of information; in PaLM's case, the tokens were gleaned from English and multilingual web documents, books and code. Minerva was the result of fine-tuning PaLM on tens of billions of tokens from scientific papers and mathematics-related web pages.

Minerva can answer prompts such as: what is the largest multiple of 30 that is less than 520? The LLM appears to be thinking through the steps, and yet all it is doing is turning the questions into a sequence of tokens, generating a statistically plausible next token, appending it to the original sequence, generating another token, and so on: a process called inference.

Google researchers fine-tuned three sizes of Minerva using underlying pre-trained PaLM models of 8 billion, 62 billion and 540 billion parameters. Minerva's performance improved with scale. On the overall MATH data set, the smallest model had 25% accuracy, the medium-sized model reached 43% and the largest breached the 50% mark.

> **"THE BIGGER MODELS KEEP DOING BETTER AND BETTER."**

The biggest model also used the least amount of fine-tuning data – it was fine-tuned on only 26 billion tokens, whereas the smallest model looked at 164 billion tokens. But the biggest model took a month to fine-tune, on specialized hardware that had eight times as much computing capacity as used for the smallest model, which was fine-tuned for only two weeks. Ideally, the biggest model would have been fine-tuned on more tokens, says Ethan Dyer at Google Research in Mountain View, California, who is a member of the Minerva team; this could have led to an even better performance. But the team felt that the computational expense wasn't feasible.

Scaling laws

That the biggest Minerva model did best was in line with studies that have revealed scaling laws – rules that govern how performance improves with model size. A study in 2020 showed that models did better when given one of three things: more parameters, more training data or more 'compute' (the number of computing operations executed during training)[4]. Performance scaled according to a power law, meaning that it improved as some power of, for example, the number of parameters.

However, researchers don't exactly know why. "The laws are purely empirical," says Irina Rish, a computer scientist at the University of Montreal and Mila – Quebec Artificial Intelligence Institute in Montreal, Canada.

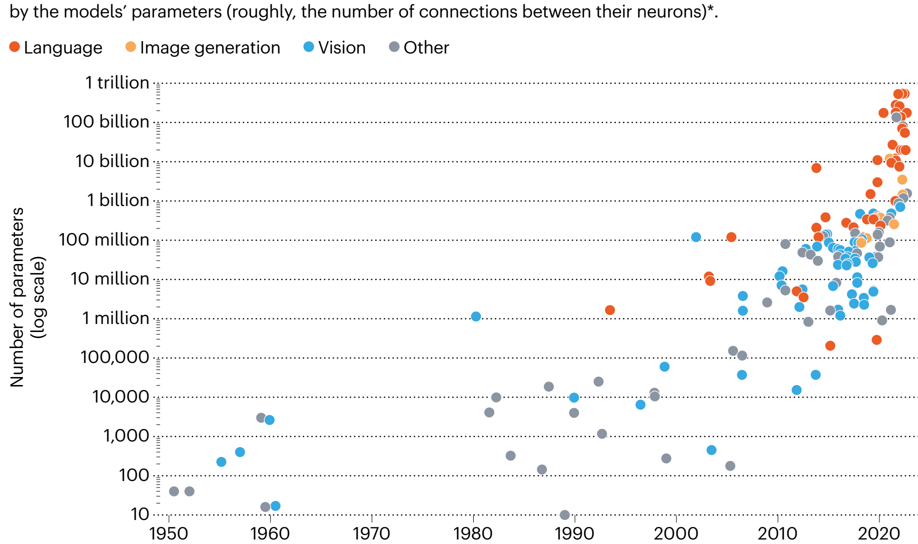
For the best results, the 2020 study suggested that as training data is doubled, model size should increase five times. Work last year slightly revised this. In March, the London-based AI firm DeepMind argued that it's best to scale up model size and training data together, and that smaller models trained on more data do better than bigger models trained on fewer data[5] (see 'Different routes to scale'). For example, DeepMind's Chinchilla model has 70 billion parameters, and was trained on 1.4 trillion tokens, whereas its 280-billion-parameter Gopher model, was trained on 300 billion tokens. Chinchilla outperforms Gopher on tasks designed to evaluate what the LLM has learnt.

Scientists at Meta Research built on this concept in February with their own small-parameter model called LLaMA, trained on up to 1.4 trillion tokens. The 13-billion-parameter version of LLaMA outperformed ChatGPT's forerunner GPT-3 (175 billion parameters), the researchers say, whereas the 65-billion-parameter version was competitive with Chinchilla and even PaLM (see go.nature.com/3kje2fj).

Last October, Ethan Caballero at McGill University in Montreal, together with Rish and others, reported finding more complex relationships between size and performance[6]. In some instances, multiple power laws can

THE DRIVE TO BIGGER AI MODELS

The scale of artificial-intelligence neural networks is growing exponentially, as measured by the models' parameters (roughly, the number of connections between their neurons)*.



*'Sparse' models, which have more than one trillion parameters but use only a fraction of them in each computation, are not shown.

ADAPTED FROM OUR WORLD IN DATA, AND FROM J. SEVILLA *ET AL.* PREPRINT AT ARXIV HTTPS://DOI.ORG/10.48550/ARXIV.2202.05924 (2022).

© 2023 Springer Nature Limited. All rights reserved.

govern how performance scales with model size, the researchers say.

For instance, in one hypothetical scenario fitting a general equation that they found, performance improves first gradually and then more rapidly with a model's size, but then dips slightly as the number of parameters continues to rise, before increasing again. The characteristics of this complex relationship are dictated by the specifics of each model and how it is trained. Ultimately, the researchers hope to be able to predict this in advance as any particular LLM is scaled up.

A separate theoretical finding also supports the drive for bigger models – a 'law of robustness' for machine learning, introduced in 2021 by Bubeck and his colleagues[7]. A model is robust if its answers remain consistent despite small perturbations in its inputs. Some AIs are notoriously fragile. If trained to recognize images of dogs, for instance, they will misclassify a test image when it's modified by a small amount of noise that wouldn't fool a person.

The more robust the AI, the better it generalizes to unseen data. Bubeck and his colleagues have shown mathematically that increasing the number of parameters in a model increases robustness, and hence ability to generalize. The law proves that scaling up is necessary for generalization, but not that it's sufficient, says Bubeck. Nonetheless, it is being used to justify the move towards bigger models, he says. "Which I think is a reasonable thing."

Minerva also took advantage of a key innovation called chain-of-thought prompting. The user prefixes their question with text that includes a couple of examples of questions and solutions, including the reasoning – illustrating a typical chain of thought – that led to the answers. During inference, the LLM takes its cues from this context and provides a step-by-step answer that looks surprisingly like reasoning. This doesn't require updates to the model's parameters, and so doesn't involve the additional computing power that fine-tuning needs.

The ability to respond to chain-of-thought prompts shows up only in LLMs with more than about 100 billion parameters. Such discoveries have helped bigger models to improve in accordance with empirical scaling laws, says Blaise Agüera y Arcas at Google Research in Seattle, Washington. "The bigger models keep doing better and better."

Reasonable concerns

François Chollet, an AI researcher at Google in Mountain View, is among the sceptics who argue that no matter how big LLMs become, they will never get near to having the ability to reason (or mimic reasoning) well enough to solve new problems reliably. An LLM only appears to reason by using templates that it has encountered before, whether in the training data or in the prompt, he says. "It cannot, on the fly, make sense of something that it has not seen before."

The best that LLMs might be able to do is to slurp in so much training data that the statistical patterns of language alone allow them to respond to questions with answers that are very close to what they've already seen.

Agüera y Arcas, however, argues that LLMs do seem to have gained some surprising abilities that they weren't trained on specifically. In particular, he points to tests designed to show whether a person has what's called theory of mind – being able to theorize or gauge the mental states of others. Take this simple example. Alice puts her glasses away in a drawer. Then Bob, unbeknown to Alice, hides the glasses under a cushion. Where will Alice look for her glasses first? A child asked this question is being tested on whether they understand that Alice has her own beliefs that might not agree with what the child knows.

In his experiments with another of Google's LLMs called Language Model for Dialogue Applications (LaMDA), Agüera y Arcas found that LaMDA responded correctly in more extended conversations of this type. To him, this was suggestive of an LLM's ability to internally model the intentions of others. "These models that are doing nothing but predicting sequences develop an extraordinary range of capabilities, including theory of mind," says Agüera y Arcas. But he concedes that these models are error-prone, and he, too, is unsure whether scaling alone, although it seems necessary, is sufficient for reliable reasoning.

> "EVERY BIG TECH COMPANY WILL NOW ATTEMPT TO DEPLOY LLMS."

Even when LLMs get the answers right, however, there is no understanding involved, says Chollet. "When you try to probe it a little bit, it becomes immediately obvious that it's all empty. ChatGPT has no model of what it is talking about," he says. "You're watching a puppet show and believing the puppets are alive."

So far, LLMs still make absurd mistakes that humans never would, says Melanie Mitchell, who studies conceptual abstraction and analogy-making in AI systems at the Santa Fe Institute in New Mexico. This has contributed to the many concerns about the safety of unleashing LLMs without guardrails into society.

An issue with the debate over whether LLMs can ever tackle genuinely new, unseen problems is that we have no way of comprehensively testing for this ability, Mitchell adds. "The current benchmarks we have are not adequate," she says. "They're not probing things systematically. We don't really know yet how to do that."

The problems of scale

While the debate plays out, there are already pressing concerns over the trend towards larger language models. One is that the data sets, computing power and expense involved in training big LLMs restricts their development – and therefore their research direction – to companies with immense computing resources. OpenAI has not confirmed the costs of creating ChatGPT, but others have estimated on the basis of the compute involved that pre-training GPT-3 (a predecessor of ChatGPT) would have cost more than US$4 million. It's probably costing OpenAI millions of dollars each month to run ChatGPT, because of the number of queries that the free chatbot is now servicing. "We are already deep into this regime," says Bubeck. "There are just a few companies that have models bigger than 100 billion parameters."

Governments are starting to step in with support that might expand the playing field. In June last year, an international team of around 1,000 academic volunteers, with funding from the French government, a US AI company called Hugging Face and others, trained a model called BLOOM with about 175 billion parameters, using $7 million worth of computing time[8]. And in November, the US Department of Energy awarded supercomputing time to a project from Rish and her colleagues, to build large models to study their behaviour. "We hope to train a Chinchilla-like 70-billion-parameter model – not necessarily the largest, but rather the one whose performance scales more effectively," says Rish.

Regardless of who gets to build them, LLMs also raise concerns about electricity consumption. For example, Google reported that training PaLM took about 3.4 gigawatt-hours over about two months. That's the annual energy consumption of about 300 US households. Google trained PaLM at its Oklahoma data centre, which it said operated on 89% carbon-free energy, being powered mostly by wind and other renewable sources. But a survey of industry AI models has shown that the majority are trained using electricity grids that are still largely powered by fossil fuels[9].

Chollet's concern is that as multiple firms begin training and using bigger models, they could start to suck up more electricity. "Every big tech company will now attempt to deploy LLMs into their products, regardless of whether it's a good idea or not," he says.

Smarter and smaller?

For many scientists, then, there's a pressing need to reduce LLM's energy consumption – to make neural networks smaller and more efficient, as well as, perhaps, smarter. Besides

© 2023 Springer Nature Limited. All rights reserved.

the energy costs of training LLMs (which, although substantial, are a one-off), the energy needed for inference – in which LLMs answer queries – can shoot up as the number of users increases. Big tech firms haven't commented on the usage costs of their models. Hugging Face, however, revealed that when its BLOOM model was deployed on the Google Cloud Platform for 18 days, in which time it answered 230,768 queries (many fewer than ChatGPT, which hit 100 million active users a month in February), it consumed, on average, 1,664 watts[10].

For comparison, our own brains are much more complicated and larger than any LLM, with 86 billion neurons and some 100 trillion synaptic connections. And yet, the human brain consumes somewhere between 20 and 50 watts of power, says Friedemann Zenke at the Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland.

So some researchers hope that mimicking aspects of the brain will help LLMs and other neural networks to become smaller, smarter and more efficient.

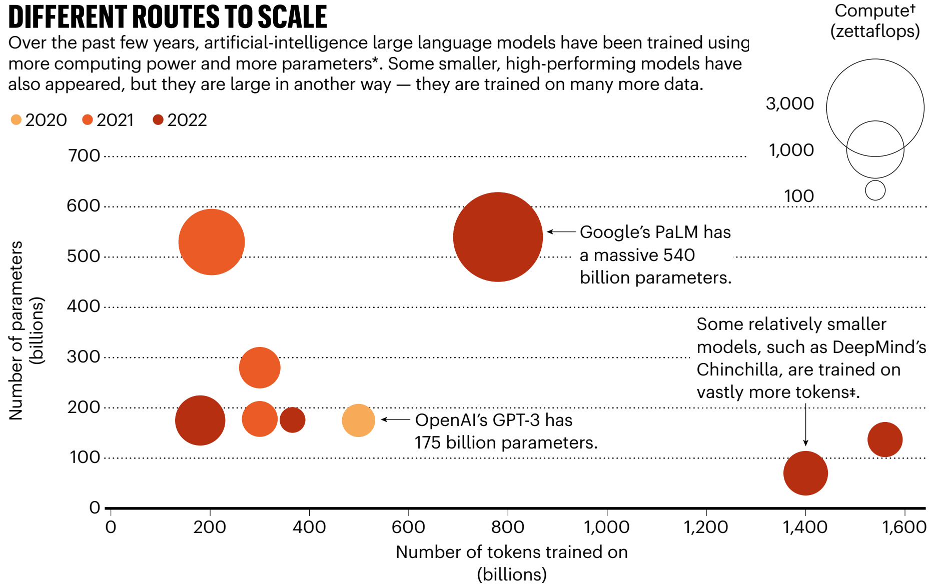
One source of the brain's overall intelligence and efficiency might be its recurrent, or feedback, connections. LLMs are, essentially, 'feedforward' networks. This means that information flows one way: from the input, through the layers of the LLM, to its output. The brain is wired differently. For example, in the human visual system, neurons connect regions of the brain that first receive visual information to areas further back. But there are also feedback connections that allow information transfer between neurons in the reverse direction. "There's maybe ten times as many feedback connections as there are feedforward connections in the [human] visual system," says Mitchell, but an LLM has no feedback connections.

Artificial neural networks that incorporate both feedforward and feedback connections are generically called recurrent neural networks (RNNs). Such networks (unlike feedforward LLMs) can discern patterns in data that change over time. That's "fundamental to how all natural intelligences experience the world and learn", says Kanaka Rajan, a computational neuroscientist at the Icahn School of Medicine at Mount Sinai in New York City. But RNNs come with challenges, says Rajan. For instance, they are hard and slow to train, making it difficult to scale them to the size of current LLMs.

Another reason brains are efficient is that biological neurons mostly remain quiet – they have only an occasional spike in activity. By contrast, the artificial neurons in most neural networks are modelled as being constantly on. Researchers are studying artificial neurons that spike (mimicking real ones), but it is difficult to adapt algorithms that train standard neural networks to networks that use spiking neurons. Still, research using small data sets (for example, 10,000 audio recordings used

DIFFERENT ROUTES TO SCALE

Over the past few years, artificial-intelligence large language models have been trained using more computing power and more parameters*. Some smaller, high-performing models have also appeared, but they are large in another way — they are trained on many more data.



*Parameters: roughly, the number of connections between neurons. †Compute: number of computing operations executed during training, measured as floating point operations (flops). ‡Tokens: words, digits or other units of information that models are trained on.

ADAPTED FROM OUR WORLD IN DATA, AND FROM J. SEVILLA *ET AL.* PREPRINT AT ARXV HTTPS://DOI.ORG/10.48550/ARXIV.2202.05924 (2022).

to train a network to recognize spoken digits) has shown that RNNs with spiking neurons outperform those with standard neurons, and, in theory, are three orders of magnitude more computationally efficient[11]. "Progress is rapid and impressive," says Sander Bohté, at Amsterdam's National Research Institute for Mathematics and Computer Science in the Netherlands (CWI), who works in this area.

As long as such spiking networks are only simulated in software, however, they cannot provide real efficiency gains (since the hardware simulating them still consumes power). Such computing elements will need to be built into hardware, on neuromorphic chips, to realize their benefits.

Energy-efficient LLMs

Meanwhile, researchers are experimenting with different ways to make existing LLMs more energy efficient, and smarter. In December 2021, DeepMind reported a system called RETRO, which combines an LLM with an external database. The LLM uses relevant text retrieved from this database during inference to help it make predictions. DeepMind's researchers showed that a 7.5-billion parameter LLM, coupled with a database of 2 trillion tokens, outperforms LLMs with 25 times more parameters[12]. The researchers wrote that this was a "more efficient approach than raw parameter scaling as we seek to build more powerful language models".

In the same month, scientists at Google Research reported another way to increase energy efficiency at scale. Their Generalist Language Model, or GLaM, has 1.2 trillion parameters[13]. But these parameters don't represent one giant neural network; internally, they are distributed between 64 smaller neural networks, alongside other layers. The LLM is trained such that during inference, it uses only two of its networks to complete a task; overall, the LLM uses only about 8% of its trillion-plus total parameters for inference, per token. According to Google, GLaM used the same amount of computing resources as were needed to train GPT-3, but consumed only about one-third of the power, because of improvements in training software and hardware. During inference, GLaM used half the computing resources that GPT-3 needed. And it outperformed GPT-3 when trained on the same amount of data.

To improve further, however, even these more energy-efficient LLMs seem destined to become bigger, using up more data and compute. Researchers will be watching to see what new behaviours emerge with scale. "Whether it will fully unlock reasoning, I'm not sure," says Bubeck. "Nobody knows."

Anil Ananthaswamy is a freelance science writer based in California.

1. Frieder, S. *et al*. Preprint at https://arxiv.org/abs/2301.13867 (2023).
2. Lewkowycz, A. *et al*. Preprint at https://arxiv.org/abs/2206.14858 (2022).
3. Chowdhery, A. *et al*. Preprint at https://arxiv.org/abs/2204.02311 (2022).
4. Kaplan, J. *et al*. Preprint at https://arxiv.org/abs/2001.08361 (2020).
5. Hoffmann, J. *et al*. Preprint at https://arxiv.org/abs/2203.15556 (2022).
6. Caballero, E. *et al*. Preprint at https://arxiv.org/abs/2210.14891 (2022).
7. Bubeck, S. *et al*. Preprint at https://arxiv.org/abs/2105.12806 (2021).
8. Le Scao, T. *et al*. Preprint at https://arxiv.org/abs/2211.05100 (2022).
9. Luccioni, A. S. & Hernandez-Garcia, A. Preprint at https://arxiv.org/abs/2302.08476 (2023).
10. Luccioni, A. S., Viguier, S. & Ligozat, A.-L. Preprint at https://arxiv.org/abs/2211.02001 (2022).
11. Yin, B. *et al*. *Nature Mach. Intell.* **3**, 905–913 (2021).
12. Borgeaud, S. *et al*. Preprint at https://arxiv.org/abs/2112.04426 (2021).
13. Du, N. *et al*. Preprint at https://arxiv.org/abs/2112.06905 (2021).

© 2023 Springer Nature Limited. All rights reserved.



CHRISTOPHE KETELS/ALAMY

Dance clubs and other indoor spaces in Belgium will soon post information about air quality.

DISEASES IN THE ROOM

The COVID pandemic has brought attention to the importance of healthy indoor air, and could spur countries to make lasting improvements. **By Dyani Lewis**

Bars in Belgium could be among the healthiest places to have a drink, come July. That's when a new law goes into effect, requiring public venues to meet air-quality targets and display real-time measurements of carbon dioxide concentrations – a proxy for how much clean air is piped in.

Consumers in Belgium will get even more information in 2025, when gyms, restaurants and indoor workspaces must all show air-quality ratings given through a certification system. In the event of a future pandemic, Belgium's rating system could determine whether or not a venue is closed.

The law, enacted in July 2022, is the boldest in a string of moves that countries have taken in the wake of the COVID-19 pandemic to make indoor spaces safer in the face of infectious diseases caused by viruses such as SARS-CoV-2 and influenza.

In March 2022, the US government launched a Clean Air in Buildings Challenge to spur building owners and operators to improve their ventilation and indoor air quality. In October last year, the state of California passed a law requiring all school buildings to provide clean indoor air. And in December, the White House announced that all federal buildings – some 1,500 in total – would meet minimum air-safety requirements. Also in December, the American Society of Heating, Refrigerating and Air Conditioning Engineers (ASHRAE) – a construction-industry body whose recommendations are adopted into law through local building codes in the United States and elsewhere – announced that it would be developing standards that take infection risk into account by June 2023.

Last June, the United Kingdom's leading engineering bodies released a report, commissioned by the government, that called for enforceable clean-air regulations to make buildings safe over their entire lifetimes (see go.nature.com/3kgsmjt). Other countries are also taking steps – for example, by deploying air-quality monitors in classrooms.

Specialists in indoor air quality are buoyed by the prospect that the pandemic could bring lasting improvements to the air we breathe indoors. The SARS-CoV-2 virus that causes COVID-19 is spread mainly in indoor spaces, as are the pathogens that lead to other infectious diseases, such as chicken pox, measles, tuberculosis and seasonal influenza.

"There's never been, in history, so much

© 2023 Springer Nature Limited. All rights reserved.

action about indoor air quality," says Lidia Morawska, an aerosol scientist at the Queensland University of Technology in Brisbane, Australia.

But huge challenges lie ahead, particularly for the existing stock of schools, office buildings and other public venues. Retrofitting them with the technology to deliver clean air at sufficient levels will be an immense – and costly – undertaking, say experts in this field. But, they argue, the benefits would outweigh the costs. By one estimate, pandemic and seasonal influenza outbreaks cost the United Kingdom £23 billion (US$27 billion) per year, on average (see 'The high cost of outbreaks'), and the country could save £174 billion over a 60-year period by improving ventilation in buildings (see go.nature.com/3ktumeg).

Making indoor spaces safe from infection could also reduce exposure to pollutants such as fine particulates from wildfire smoke and cooking, volatile organic compounds leached from furniture, and allergy-causing moulds and pollen. But it could also raise energy costs and contribute to greenhouse-gas emissions.

Researchers are still working to pin down how best to ventilate indoor spaces to prevent infections from spreading, and what alternative technologies might replace or enhance mechanical ventilation systems. But many say that enough is already known to start demanding safer indoor spaces.

It's a race against time. As concern over COVID-19 wanes, experts wonder how much progress countries will make before the next big outbreak of an airborne infectious disease.

HOW MUCH CLEAN AIR IS ENOUGH?

A task force of researchers proposed ventilation rates for buildings using several metrics*, with the aim of reducing the risks of transmission of airborne respiratory diseases.

Assessment	Equivalent air exchanges per hour	Cubic feet per minute per person	Litres per second per person
Good	4	21	10
Better	6	30	14
Best	>6	>30	>14

*Not shown: volumetric flow rates per floor area.

SOURCE: REF. 6

Reducing infections

When COVID-19 reached pandemic status in early 2020, health officials didn't pay much attention to the risks of indoor air. Initially, the World Health Organization (WHO) dismissed the role of airborne transmission and focused – incorrectly – on transmission through contaminated surfaces. But even when public-health authorities began recommending better ventilation as a way of preventing infection, they offered only vague guidance. Authorities told people to open windows and bring in as much outdoor air as possible with mechanical ventilation systems, without giving specific numbers.

Such advice sowed confusion, says Joseph Allen, a building hygienist at the Harvard T.H. Chan School of Public Health in Boston, Massachusetts. "You can't tell people to bring in more outdoor air without answering how much," he says.

Allen was one of the first to put a value on how much ventilation people should be aiming for. In June 2020, he and his colleagues recommended that schools wanting to reopen their doors after lockdowns should deliver four to six air changes per hour to their classrooms[1] – changes in which the entire volume of air in the room is replaced. That amounts to a ventilation rate of 10–14 litres per second per person. Most schools were achieving much less than that, however. A study of California classrooms, for example, found that most failed to meet that level of ventilation[2]. The WHO issued its own guidelines in March 2021, recommending a ventilation rate of 10 litres per second per person outside health-care settings.

> **"You can't tell people to bring in more outdoor air without answering how much."**

In theory, the pandemic provided the perfect opportunity to gather real-world data to see whether low ventilation rates were associated with outbreaks, and to test different rates of ventilation to see which resulted in reduced infection rates. But health officials only rarely considered ventilation when investigating major outbreaks of COVID-19. Yuguo Li, a mechanical engineer at the University of Hong Kong, estimates that fewer than ten investigations measured ventilation rates in venues where outbreaks occurred, because airborne transmission was not on people's radar.

Instead, researchers tried to gain clues through observational studies. Morawska was involved in one that looked at 10,000 school classrooms in the Marche region of Italy. In the 316 classrooms that had mechanical ventilation with rates of 1.4–14 litres per second per person, the students' risk of infection was reduced by at least 74% over a 4-month period at the end of 2021, compared with that for students in classrooms that relied on windows for ventilation. This group typically received less than 1 litre per second per person. When ventilation rates were at least 10 litres per second per student, the infection risk was 80% lower[3].

Evidence is also growing about other technologies that remove infectious particles from the air. One study[4] explored the effectiveness of two air cleaners fitted with high-efficiency particulate absorbing (HEPA) filters, placed in a 54-square-metre conference room with a dummy that generated aerosol particles similar to those that transmit SARS-CoV-2. The cleaners reduced the aerosol exposure of three dummy participants by 65%. That's just shy of the 72% reduction achieved by masking all of the dummy participants[4].

Another study, by civil engineer Bert Blocken at the Catholic University of Leuven (KU Leuven) in Belgium, found that ventilation combined with air cleaning, equivalent to 6 air changes per hour in total, reduced exhaled aerosol concentrations in a gym to 5–10% of what they would have been without these measures[5]. That concentration substantially reduces infection risk, says Blocken. He adds that air cleaners are an underappreciated technology that could be readily deployed in buildings that don't have mechanical ventilation systems capable of providing enough clean air, or where operating such systems would consume too much energy. The state of Victoria in Australia took this approach, distributing portable air cleaners to all of its 110,000 classrooms in 2022.

Last November, the *Lancet* COVID-19 Commission's Task Force on Safe Work, Safe School, and Safe Travel, chaired by Allen, published concrete guidelines for clean-air delivery rates – using ventilation, air filtration or other means – to reduce airborne infections[6]. To achieve what the report describes as the 'best' air quality, it recommends more than 6 air changes per hour, or 14 litres per second per person (see 'How much clean air is enough?').

Legal limits

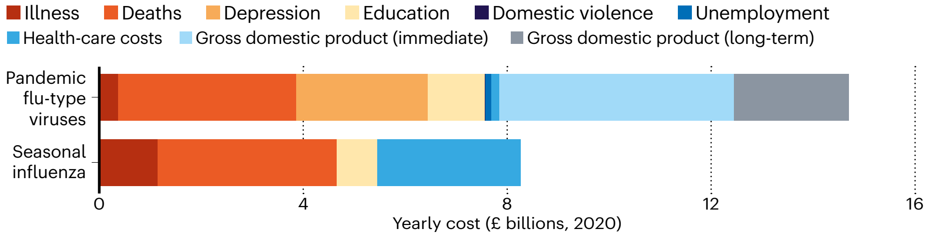
Ventilation requirements can be complicated, because they change depending on how big the space is, how many people are in it and how active they are. So some researchers advocate using a shortcut – setting maximum carbon dioxide concentrations. CO_2 is frequently used as a proxy measure for ventilation and indoor air quality[7]. Because people exhale CO_2 as they breathe, levels of the gas can shoot up if a space is crowded or if there is insufficient ventilation to replace the exhaled air – which might contain infectious viruses – with clean air.

Until 1999, ASHRAE standards included a recommended limit for CO_2 of 1,000 parts per million (p.p.m.). At this concentration, according to research conducted in the 1930s, building occupants' perception of

© 2023 Springer Nature Limited. All rights reserved.

THE HIGH COST OF OUTBREAKS

A study estimated the annual costs that influenza would exact on the United Kingdom over a 60-year period. Outbreaks of pandemic respiratory viruses, which include flu and coronaviruses such as SARS-CoV-2, are nearly twice as costly as seasonal flu outbreaks.



SOURCE: NERA ECONOMIC CONSULTING

body odour would be kept at an acceptable level. Since then, research has shown that when concentrations exceed 1,000 p.p.m., CO_2 can cause drowsiness and can impair cognitive performance on decision-making and problem-solving tasks[8].

A small study published in September 2022 – and yet to be peer reviewed – directly connected CO_2 levels with those of infectious pathogens. The authors tested air samples in nurseries, schools, universities and care homes for the presence of respiratory pathogens. Rooms that had higher CO_2 levels were associated with higher levels of respiratory pathogens[9].

In August 2021, the UK government began distributing CO_2 sensors to all school classrooms so that teachers could use the devices to decide when to open windows or increase ventilation. Similar schemes have been rolled out in Europe, the United States and elsewhere, although none has yet been evaluated for its ability to reduce infection rates.

Relying on CO_2 readings has drawbacks, however. Concentrations can creep up even when the infection risk remains low, such as when using portable air cleaners – which do not remove CO_2 from the air – or when cooking. CO_2 is useful, says chemist Nicola Carslaw at the University of York, UK, who studies indoor-air pollutants, "but it's definitely not the whole story".

Despite these issues, Morawska says that CO_2 monitors should be widely deployed as an inexpensive, readily available tool that could be installed in every indoor space, much like smoke alarms. But displaying CO_2 read-outs on its own is not enough, she adds, because it places the onus on room occupants to track air quality and decide what to do if readings are high.

Morawska would also like to see laws that mandate maximum CO_2 levels permissible in public buildings, so that the responsibility is placed back on building operators and government regulators. A handful of governments have already done just that. Last year, Morawska and her colleague Wei Huang at Peking University in Beijing reviewed air-quality laws in more than 100 countries. Only 12 had national standards for indoor air quality that specified threshold limits for pollutants. And only 8 of those – including China, South Korea, India, Poland and Hungary – set limits for CO_2 concentration, most between 800 p.p.m. and 1,000 p.p.m. (ref. 10).

Japan has had a law to regulate indoor air quality since 1970, which mandates that buildings must not exceed indoor CO_2 concentrations of 1,000 p.p.m.. The law requires that building managers assess air quality every two months, report results to the government and establish remediation plans if the air quality does not meet the standards. But almost 30% of buildings exceeded the CO_2 limit in 2017, according to a 2020 report[11].

Still, the Japanese laws work, says Kazukiyo Kumagai, a public-health engineer at the California Department of Health in Richmond. "Japan is in a better condition" than the United States when it comes to indoor air quality, he says. Adopting a Japanese-style approach of regular monitoring and reporting might work elsewhere, he adds.

Legal limits could become more common. The new Belgian law, for example, comes into effect in July this year and stipulates that public venues ventilate at a rate of 40 cubic metres per hour so that CO_2 does not exceed 900 p.p.m.. If air filtration is used, a lower ventilation rate of 25 cubic metres per hour is enough, and CO_2 can reach a maximum level of 1,200 p.p.m..

Legislating indoor air quality is "tricky" says Catherine Noakes, a mechanical engineer at the University of Leeds, UK, who contributed to that country's report into infection-resilient buildings. "One of the challenges with indoor air," she says, "is who owns it?" The responsibility can be distributed across government departments and agencies, depending on how the building is used. A school's indoor air might be the responsibility of the education department, whereas office buildings could be regulated by an occupational health and safety agency.

That's the situation in the United States, where no agency currently has the authority to regulate indoor air, says Andrew Persily, a mechanical engineer at the National Institute of Standards and Technology in Gaithersburg, Maryland. In Belgium, too, the new national law doesn't cover schools, which are the responsibility of regional governments. And in Japan, a separate law for school buildings specifies a higher CO_2 limit of 1,500 p.p.m., a level many regard as too high.

Setting standards

In the absence of national laws, professional bodies that set air-quality standards are starting to act. When ASHRAE releases its infection-mitigation standard in June, the hope is that these recommended targets will be adopted into local building codes that new buildings must comply with.

"We have always addressed indoor air quality, but not specifically for pathogen mitigation," says engineer Ginger Scoggins, the president-elect of ASHRAE, who is based in North Carolina. ASHRAE could face some pushback. Scoggins says that when the society made a previous change to increase the ventilation requirement from 5 cubic feet per minute to 15 (2.4 litres per second to 7.1 litres per second), many people in the warm parts of the United States were angry because it would drive up energy costs from air conditioning. Her local school board passed a ruling that its classrooms only needed to get to 7.5.

Even though ASHRAE standards are not enforced, they will make a difference, says Allen. Aside from influencing how buildings are constructed, more stringent ASHRAE standards send a strong signal to businesses in older buildings about what the gold standard for indoor air quality looks like.

An economic case could be made for better indoor air, says Noakes. The cost–benefit analysis conducted for the UK report found that the country could save £3 billion per year over a 60-year period by improving ventilation.

Researchers say it will take time to lower the infection risks inside buildings. "We are looking at 30 years," says Morawska. "But we are talking about the future of our society."

Dyani Lewis is a senior reporter with *Nature* in Melbourne, Australia.

1. Jones, E. *et al. Schools for Health: Risk Reduction Strategies for Reopening Schools* (Harvard T.H. Chan School of Public Health Healthy Buildings Program, 2020).
2. Mendell, M. J. *et al. Indoor Air* **23**, 515–528 (2013).
3. Buonanno, G. *et al. Front. Public Health* **10**, 1087087 (2022).
4. Lindsley, W. G. *et al. Morb. Mortal. Wkly Rep.* **70**, 972–976 (2021).
5. Blocken, B. *et al. Build. Environ.* **193**, 107659 (2021).
6. The *Lancet* COVID-19 Commission Task Force on Safe Work, Safe School, and Safe Travel. *Proposed Non-infectious Air Delivery Rates (NADR) for Reducing Exposure to Airborne Respiratory Infectious Diseases* (*Lancet* COVID-19 Commission, 2022).
7. Peng, Z. & Jimenez, J. L. *Environ. Sci. Technol. Lett.* **8**, 392–397 (2021).
8. Wargocki, P., Porras-Salazar, J. A., Contreras-Espinoza, S. & Bahnfleth, W. *Build. Environ.* **173**, 106749 (2020).
9. Raymenants, J. *et al.* Preprint at medRxiv https://doi.org/10.1101/2022.09.23.22280263 (2022).
10. Morawska, L. & Huang, W. In *Handbook of Indoor Air Quality* (eds Zhang, Y. *et al.*) (Springer, 2022).
11. Hayashi, M., Kobayashi, K., Kim, H. & Kaihara, N. *J. Natl Inst. Public Health* **69**, 63–72 (2020).

© 2023 Springer Nature Limited. All rights reserved.

Books & arts



GETTY

Centuries of development led to the bicycle's spoked wheel — which brought mobility and freedom for many.

Seven objects that made the modern world

Big technological advances rest on small things, a structural engineer explains. **By Anna Novitzky**

Ten thousand years ago, the most advanced tools were made of chipped stone. Today, much technology is so complicated that it's almost "indistinguishable from magic", to borrow Arthur C. Clarke's phrase. How did we get here? In *Nuts and Bolts*, structural engineer Roma Agrawal investigates the story of basic technological developments, and shows how intimately entwined they are with humanity's own history.

She contends that the modern world has its foundations in seven humble inventions: the nail, wheel, spring, magnet, lens, string and pump. Building on her lifelong fascination with opening things up to see what makes them work, she explores the science of each of these mini 'machines', and follows their history from ancient beginnings to manifestations of modern engineering, small and large.

The spring, for example, she characterizes as "humanity's first tool that allowed us to store energy and then release it when we wanted". She charts its development from bows and arrows to the vast steel coils that help skyscrapers to withstand earthquakes, and the silicon hairsprings that maintain the accuracy of the most exclusive mechanical watches.

With a clear, lively and engaging style – and many puns – Agrawal encourages a new perspective on the inventions that keep the world rolling. The wheel, she explains, was first used not for transport, but to make pottery. It wasn't for another 700 years that it was turned on its side and attached to an axle; the earliest surviving wheeled vehicles are from around 3200 BC, in what's now Russia. A multitude of refinements followed, technological advances ushering in sweeping social change. Spoked wheels supplanted solid ones, their lightness enabling fast travel and improved trade. Wheels with wire spokes led to the bicycle, a source of freedom for many who couldn't afford carriages or cars.

Gear change

Serrated wheels became gears, tools for changing the direction and magnitude of forces that prompted advances of the Industrial Revolution and beyond. In the late nineteenth century, for instance, Josephine Cochran used wheels and gears to develop the first commercially viable dishwasher – one of myriad appliances that would cut the time spent on housework and free many women to enter the workforce. And wheels are ancestors of the gyroscopes that steer the International Space Station. As Agrawal says, 5,000 years of change have been driven by constant reinvention of the wheel.

Refreshingly, this world-trotting tale focuses on "the often hidden or unacknowledged contribution of minoritised people".



Nuts and Bolts: Seven Small Inventions That Changed the World (in a Big Way)
Roma Agrawal
Hodder & Stoughton (2023)

© 2023 Springer Nature Limited. All rights reserved.

Soviet women work as fighter pilots in the Second World War; spinning wheels take a key role in the Indian independence movement. Even familiar names have an unexpected spin. Telephone inventor Alexander Graham Bell, for example, originally set out to convert the vibrations of speech into something that his mother and wife, both deaf, could see.

The global tapestry is shot through with Agrawal's personal stories. She makes her own nails at a blacksmith's forge, and describes movingly how she owes the birth of her daughter to the lens, without which the microscopes that were crucial to the development of *in vitro* fertilization (IVF) would not have existed. Her family history illustrates how using magnets to modulate electric pulses enabled technologies that have repeatedly revolutionized global communications. Her uncle, looking for work in Italy in the 1960s, sends telegrams to his family in Mumbai, India. An aunt in Connecticut in the 1970s books a trunk call across continents through a network of telephone exchanges. Agrawal's daughter speaks her early words to her own grandmother by video call during the COVID-19 pandemic.

All that glitters

The reader can't help but be swept along in Agrawal's enthusiasm, but she's at pains to point out that 'progress' isn't always good. Spoked wheels made chariots nimble enough to be used in war. Springs are crucial to guns. String led to a textile industry that bankrolled the British Empire while laying waste to the Indian economy (hence the significance of the spinning wheel). The fashion industry now generates around 10% of global carbon dioxide emissions and vast amounts of other pollution.

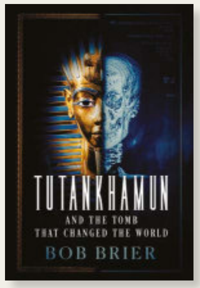
The question of engineering's impact on the planet and on society is a common thread. By drawing out stories of technology's use for good and ill, Agrawal encourages deep thinking about where we go from here.

She shows how, to manage humanity's devastation of the planet, we must demystify the 'black box' of technology. "Understanding the nuts and bolts of our objects", she says, leads us to appreciate the effort, ingenuity and raw materials that went into them, and stop seeing them as disposable. We become more likely to find ways to extend their lives, to reconstitute and repair them. Agrawal points to sustainable engineering practices, community-awareness programmes and repair movements as ways to help fend off environmental disaster. On an individual level, taking ownership of our possessions through repair brings a fulfilling "sense of happiness, satisfaction, and achievement".

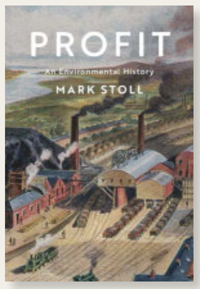
So the history of engineering tells us about who we are and who we have been. It can even lead us to who we want to be.

Anna Novitzky is a subeditor for *Nature* in London.

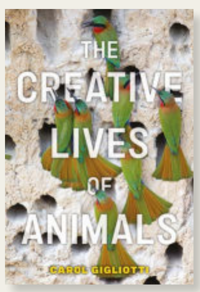
Books in brief



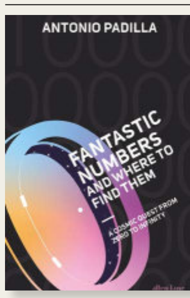
Tutankhamun and the Tomb That Changed the World
Bob Brier *Oxford Univ. Press* (2022)
Tutankhamun still raises many questions, argues Egyptologist Bob Brier in his stylish book celebrating the century since the pharaoh's tomb was rediscovered. He has been represented as a frail boy with a club foot, but he might have been an athlete who hunted, and later a warrior, says Brier. As DNA sequencing of mummies improves, Brier anticipates, Tutankhamun's parents will finally be identified, along with diseases that might have caused his death. The second 100 years of Tutankhamun research could be "even more exciting than the first".



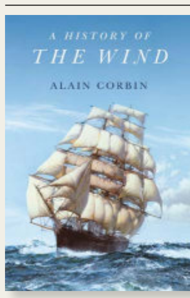
Profit
Mark Stoll *Polity* (2023)
A smartphone is a "Pandora's box of environmental evils", writes Mark Stoll. The device epitomizes both the benefits and burdens of capitalism, he argues in his history of profit from ancient times to the present, the first by an environmental historian. With knowledge, skill and stories of inventors, entrepreneurs and conservationists, he traces developments in technology, transportation, energy, communication, trade and finance. We cannot live with capitalism or without it, he says, so we must labour to "ameliorate its worst effects."



The Creative Lives of Animals
Carol Gigliotti *NYU Press* (2022)
Many songbirds are born without the ability to sing. So should those that learn — and other animals — be called creative? Carol Gigliotti interviews scientists who think they should be, and agrees with them. An animal activist, author and artist who has taught design and dynamic media, she defines creativity as a "dynamic process" in which individuals generate "novel and meaningful behaviour" that might affect others at cultural, species and evolutionary levels. Regrettably, her appealing discussion contains no images of animals.



Fantastic Numbers and Where To Find Them
Antonio Padilla *Allen Lane* (2022)
As a maths student, Antonio Padilla got a zero for a correct proof because it lacked rigour. The experience turned him into a theoretical physicist. "With mathematics, the physicist can dance, and with physics, the mathematician can sing," he says in his account of numbers that unlock the Universe, such as the cosmological constant. Although not for the mathematically faint-hearted, it sparkles — even comparing quantum theory's astonishing support of light as a particle rather than a wave with Greta Thunberg suddenly endorsing Donald Trump!



A History of the Wind
Alain Corbin (transl. William Peniston) *Polity* (2022)
For millennia, little was known about the science of the wind. Aristotle saw it as an elemental fluid, alongside water, earth and fire. In the late eighteenth century, air's chemical composition was revealed; later came the development of meteorology through mapping of global air currents and the concepts of the troposphere, stratosphere and jet stream. All this forms part of Alain Corbin's brief history, but he focuses on the reaction of artists, writers and travellers since ancient Greece to wind's "indecipherable enigma". **Andrew Robinson**

© 2023 Springer Nature Limited. All rights reserved.

Comment



AKHTAR SOOMRO/REUTERS

A family navigates flood waters to reach a village in Mehar, Pakistan, after heavy rains in August 2022.

Flash floods: why are more of them devastating the world's driest regions?

Jie Yin, Yao Gao, Ruishan Chen, Dapeng Yu, Robert Wilby, Nigel Wright, Yong Ge, Jeremy Bricker, Huili Gong & Mingfu Guan

Shifting weather, changing settlement patterns and a lack of preparedness mean that dryland areas are most at risk from flooding. Researchers need to focus on data collection, early-warning systems, flood protection and more.

Last year, around two-thirds of Pakistan was affected by widespread flash flooding, with more than 1,500 people killed and around 33 million made homeless. Almost 2,000 people died in flash floods across Africa, and parts of the United Arab Emirates, Iran, Saudi Arabia, Qatar, Oman and Yemen were inundated with water.

Flash floods are a growing threat in some of the world's driest regions. Deluges can trigger sudden and rapid torrents of run-off that flow down dry river beds and rocky channels.

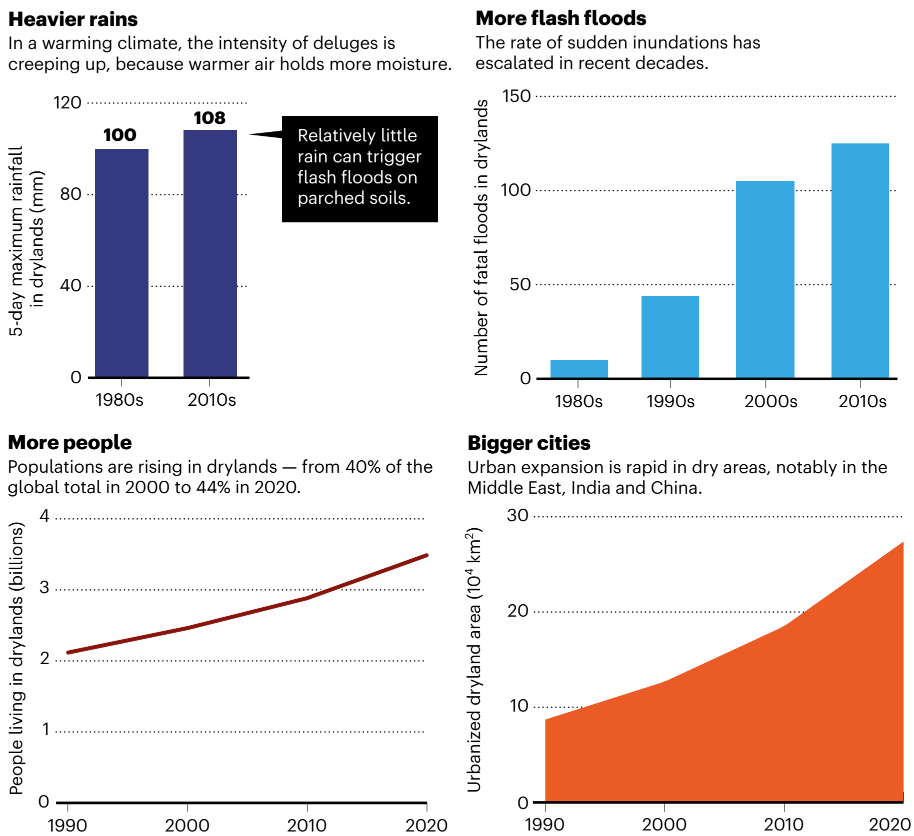
Because parched soils repel water rather than allowing it to soak in, flash floods can be more devastating in drylands than in wetter areas. Surges can result from relatively small amounts of rain, as little as 10 millimetres in one hour. By comparison, floods in wetter regions typically follow more prolonged bouts of rainfall.

Rapid run-off also erodes soils, adding sediment and debris to the water. The dangers are greatest in mountainous areas, where water can gush suddenly through gullies and down slopes. For example, a flash flood last year in Datong town in Qinghai province, China, washed away more than 1,500 homes in less than one hour.

© 2023 Springer Nature Limited. All rights reserved.

DRYLANDS: FLASH-FLOOD RISKS

The number and intensity of flash floods in arid and semi-arid areas are rising as the climate warms. Swelling cities and populations are also putting more people at risk.



SOURCE: EM-DAT, CRED/UC LOUVAIN (WWW.EMDAT.BE); ANALYSIS BY J. YIN *ET AL.*

Yet residents of drylands are underprepared for flash floods, compared with people in the tropics. In dry areas, where drought is a big concern, settlements tend to be near rivers or on floodplains. Dams and levees are low priorities.

Dryland residents are paying a high price. Our analysis using the Emergency Events Database (EM-DAT; www.emdat.be) shows that, since 2000, such regions experienced less than half (47%) of deadly flash floods globally, yet saw almost three-quarters (74%) of related deaths (see 'Global flash-flood disasters'). The majority of these floods (87%) and associated deaths (97%) occurred in low- and middle-income countries.

A slew of other factors will put many more people at risk from flash floods in future (see 'Drylands: flash-flood risks'). Climate change is making such events more intense and frequent[1–3]. In parts of Pakistan, for example, the 5-day maximum level of rainfall is 75% greater today than it was before 1900 (see go.nature.com/41awzzj). Our analysis shows that globally, the rate of dryland flash flooding was 20 times higher between 2000 and 2022 than it was between 1900 and 1999 (numbering 278 and 64 floods, respectively).

More people are living in drylands – from 40% of the global population in 2000 to 44% in 2020, by our calculation. Such regions will host 51% of the world's population growth by 2025, with half of those people living in low-income countries, and only 1% in developed ones[4]. Cities are expanding rapidly in areas such as the Middle East, India and China[5]. And the proportion of the world's land area classed as arid or semi-arid is projected to increase as the world warms – from 41% in 2000 to 48% by 2025 (ref. 4).

Pakistan is one of the most exposed nations. Almost one-third of the country's population (72 million of 230 million people) lives in high-risk flash-flood zones. As ever, vulnerable groups such as older people, women and children experience the worst effects.

Researchers, practitioners and policymakers need to model, assess and mitigate the risk of flash flooding in drylands in a changing environment. Here, we set out six research priorities for doing so.

Gather data and map the risks

Assessing the likelihood and consequences of flash floods requires a lot of data. These include: meteorological and hydrological measurements; topographic data and digital elevation models; records of assets and previous losses; and census data. However, most dryland regions lack much of this information, because national governments often don't recognize its value and many lack the resources to gather it.

Researchers, international organizations and non-governmental organizations need to help such countries to update their flood-risk information. Initiatives include the Global Flood Partnership run by the European Union (https://gfp.jrc.ec.europa.eu), which is developing flood observation and modelling infrastructure in low- to middle-income countries such as Myanmar. Governments should map flash-flood risks and distribute them online (as the United Kingdom does; see go.nature.com/3xt9hfz) to pinpoint locations that are prone to flooding, restrict construction and target flood-protection measures.

Flood risk needs to be carefully communicated, because it can exacerbate inequalities. In drylands, low-income and minority populations tend to live in flood-prone areas such as riversides[6]. If maps show these areas to be high risk, they could be deprived of local infrastructure investment, increasing their vulnerability. Socio-economic data should therefore also be conveyed to decision makers alongside the maps.

Develop early-warning systems

Early-warning systems for flash floods are operating in some wealthy countries, including the United States and in EU nations. For example, the US system (FLASH) forecasts surface water flows at 10-minute intervals and with a spatial resolution of 1 kilometre. These forecasts combine real-time observations from radar, satellites and sensors with hydrological models, and can provide a few hours' notice of flash floods. If preparation and response plans are also in place, that might give enough time for governments to allocate rescue resources and for residents to protect homes and move to safety points. However, few low- and middle-income nations have these systems, because they lack the data and models to support them, as well as political incentives and funding.

To rectify this, the World Meteorological Organization is running a programme to provide weather forecasters and

© 2023 Springer Nature Limited. All rights reserved.



FRED GREAVES/REUTERS

Vehicles cross a bridge over the Yolo Bypass in Sacramento, California, where weirs divert flood waters away from houses onto floodplains.

disaster-management agencies with real-time information around flash-flood threats (see go.nature.com/3xsw4fg). So far, 28 warning systems are up and running, with another 39 in development. Eight of the operational systems are in drylands, together with ten of those still in development. More dryland nations need to be engaged.

Forecasting of flash floods in drylands will need tailoring to local contexts, including soil types, terrains and sediment loads. Where data are scarce, uncertainties and limits need to be better understood. Use of wireless sensors and alternative sources of local data, such as use of social media and crowdsourcing, need to be explored.

Boost resilience to sudden floods

Dams, barriers and dikes will need to be installed in drylands to protect settlements. For example, China has been building thousands of kilometres of levees along the main stream and tributaries of the Yellow River. This is expensive, however, involving hundreds of projects with costs ranging from millions to billions of dollars.

Researchers need to examine the trade-offs. For example, protection measures might encourage more development in flood-prone areas — as was the case in Columbia, South Carolina, which experienced extensive damage from a flash flood in 2015. And dams and levees built to hold back small floods could offer a false sense of security from rare, extreme inundations[7], including one that flooded the Atacama region in South America in 2015. Failure of one protection element, such as a reservoir, can send flood waters cascading downstream, as was narrowly averted at California's Oroville Dam in 2017. The environmental impacts need to be understood, such as those resulting from stopping the silt and sediment that fertilize the floodplain.

"Past experience is important for recognizing risks and adapting behaviours."

Nature-based solutions are cheaper. For example, in Pakistan, the government of Khyber Pakhtunkhwa province has restored almost 350,000 hectares of forests in its 'Billion Tree Tsunami' planting drive. In the Yolo Bypass project in Sacramento, California, levees are being replaced with weirs to divert flood waters away from houses and onto natural floodplains. Large residential areas should adopt 'sponge city' techniques that reduce run-off, such as permeable pavements, green roofs, rainwater harvesting and swales (ditches). These are being installed in cities throughout China, including Xining in Qinghai province, the largest city on the Tibetan Plateau.

Researchers should aim to quantify how effective natural features are at storing and conveying water, and how resilient they are to extreme floods. Local differences in soil and vegetation conditions must be considered. More needs to be known about how the systems degrade and how to maintain them. Promising projects include the World Bank's Global Program on Nature-Based Solutions for Climate Resilience and the Sindh Resilience Project in Pakistan, as well as government-led water and soil conservation initiatives in the Loess Plateau in north-central China.

Land-use regulations need to be tightened, such as by preventing informal development in floodplains. Flood-protection standards for buildings need to be improved. For example, the International Residential Code, used as the basis of many national and

© 2023 Springer Nature Limited. All rights reserved.

local guidelines, requires the lowest floors of buildings to be 30 centimetres above the average level of a 'once a century' flood. Such levels no longer reflect flood probability in a warming world.

Plan for long-term resettlement

In flood-prone areas, governments should develop plans for temporary and even permanent resettlement. Elevated roads, evacuation routes and emergency shelters must be designed, such as those used in Indonesia for tsunamis and in the United States for hurricanes. In the longer term, as climate-change impacts accelerate, managed retreat might be required[8]. For example, the Chinese government has relocated more than 600,000 people from flash-flood zones in the arid mountains of southern Shaanxi province over the past decade.

Managed resettlement is more equitable and efficient than leaving people to fend for themselves. Cross-disciplinary research in social science, geography, engineering and psychology is required to investigate the best way to manage retreat, by providing essential services and assistance while avoiding food insecurity and lost livelihoods.

Provide finance beyond disaster aid

Risk-transfer schemes – such as flood insurance, catastrophe bonds and solidarity funds – should be developed for drylands. For example, China is developing a national system of flood insurance products. These have been incorporated into the national emergency-management system and launched in dozens of pilot cities, including Zhengzhou in Henan province. Such products include weather index insurance, which offers compensation for loss of crops and livestock when rainfall exceeds a certain threshold. Catastrophe insurance protects businesses and residences against extreme events that are too expensive for insurers to cover, such as super typhoons or floods that occur only once in 1,000 years.

Algeria has made catastrophe insurance mandatory. Every insurance company there must join its national Catastrophe Risk Insurance Pool to cover disaster losses in proportion to their market share. All properties must be insured, with premiums set by the government.

Researchers need to study the impacts of such schemes, especially on equity. For example, vulnerable families living in flood-prone zones might not be able to afford high premiums[9]. Around 50% of losses in wealthy countries are covered by insurance, but there is less than 10% coverage in many poorer countries. For example, claims for Hurricane Sandy in the United States in 2012 covered 50% of the damages. By contrast, after a record-breaking flash flood in China's Henan province in July 2021, flood insurance claims exceeded US$1.7 billion – only 11% of the total losses.

Partnerships between countries – such as China's 'Belt and Road' project and the 'Build Back Better World' initiative proposed by the G7 group of advanced economies – should be sought to establish solidarity funds that can share the risk of large disasters, as the EU has done. The United Nations climate 'loss and damage' fund, approved last year at the COP27 summit in Sharm El-Sheikh, Egypt, should be leveraged to finance long-term flood-risk reduction in drylands, once mechanisms are formulated.

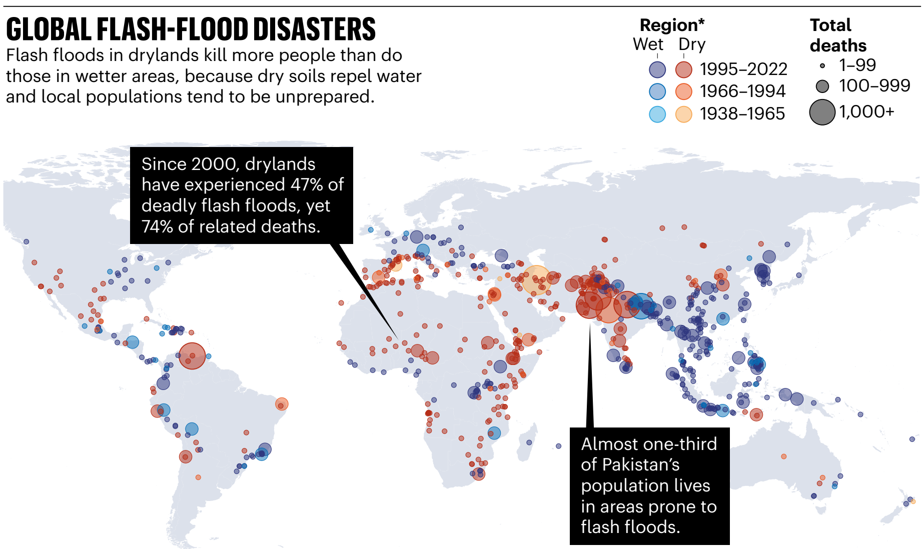
Improve public awareness and responses

People who live in drylands need to be more aware of flash-flood risks. Policymakers should offer guidance to help residents to adapt their behaviours and take precautionary measures – such as by installing flood-protection structures, knowing evacuation routes and warning signals, identifying areas prone to flooding or landslides, and avoiding walking or driving through flooded areas and standing water.

Researchers need to explore how best to communicate the effectiveness of measures, and how perceptions of risk influence responses to flash floods[10]. Social factors need to be explored – such as age, gender, education, income, ownership, culture, religion and political context. For example, past experience is important for recognizing risks and adapting behaviours. But, for rare events, memories fade and only catastrophes are remembered for a long time.

Collective behaviours also need to be understood[11]. Social-media messages can provide rapid information about locations, impacts and responses to dryland floods, as revealed through data mining and artificial intelligence. Models can integrate complex human behaviour into quantitative risk assessments, to improve management strategies. Given that flood risks are evolving, it is important to adapt policies as information becomes available. Interdisciplinary analysis will also be needed to evaluate the robustness of flood-resilience strategies for a range of risks and future scenarios[12].

Before more lives are lost, let's offer dryland residents the means to assure their safety.



GLOBAL FLASH-FLOOD DISASTERS
Flash floods in drylands kill more people than do those in wetter areas, because dry soils repel water and local populations tend to be unprepared.

*Based on the Köppen climate classification, in which regions other than 'dry' are classed as 'wet'.

SOURCE: EM-DAT, CRED/UC LOUVAIN (WWW.EMDAT.BE); ANALYSIS BY J. YIN *ET AL.*

The authors

Mingfu Guan is assistant professor of hydro-environment and geohazard in the Department of Civil Engineering, the University of Hong Kong, Hong Kong.
Jie Yin, Yao Gao, Ruishan Chen, **Dapeng Yu**, **Robert Wilby**, **Nigel Wright**, **Yong Ge**, **Jeremy Bricker**, **Huili Gong**.
A full list of author affiliations accompanies this Comment online (see go.nature.com/3mcjtvb).
e-mail: mfguan@hku.hk

1. Donat, M. G., Lowry, A. L., Alexander, L. V., O'Gorman, P. A. & Maher, N. *Nature Clim. Change* **6**, 508–513 (2016).
2. Guerreiro, S. B. *et al. Nature Clim. Change* **8**, 803–807 (2018).
3. Fowler, H. J. *et al. Nature Rev. Earth Environ.* **2**, 107–122 (2021).
4. Huang, J., Yu, H., Guan, X., Wang, G. & Guo, R. *Nature Clim. Change* **6**, 166–171 (2016).
5. Ren, Q. *et al. Nature Sustain.* **5**, 869–878 (2022).
6. Rentschler, J., Salhab, M. & Jafino, B. A. *Nature Commun.* **13**, 3527 (2022).
7. Logan, T. M., Guikema, S. D. & Bricker, J. D. *Nature Sustain.* **1**, 526–530 (2018).
8. Hino, M., Field, C. B. & Mach, K. J. *Nature Clim. Change* **7**, 364–370 (2017).
9. Jongman, B. *et al. Nature Clim. Change* **4**, 264–268 (2014).
10. Bubeck, P., Botzen, W. J. W. & Aerts, J. C. J. H. *Risk Anal.* **32**, 1481–1495 (2012).
11. Aerts, J. C. J. H. *et al. Nature Clim. Change* **8**, 193–199 (2018).
12. Dawson, R. J. *et al. Glob. Environ. Change* **21**, 628–646 (2011).

The authors declare no competing interests.

© 2023 Springer Nature Limited. All rights reserved.

Correspondence

ChatGPT: not all languages are equal

The machine-learning system ChatGPT has a remarkable capacity to generate sophisticated and intelligible text, with far-reaching implications for research development and dissemination (see *Nature* **614**, 214–216 (2023); Eva A. M. van Dis *et al.* *Nature* **614**, 224–226; 2023). However, current debate over the technology focuses on users' experiences in English. I asked ChatGPT a set of neuroscience questions in English; when I asked for answers to the same set in French and Arabic, the results were surprising.

French and Arabic are among the world's most commonly spoken languages and they have a widespread presence on the Internet. However, the richness of ChatGPT's response and the intelligibility of its writing in both languages were notably inferior to those in English. In Arabic, answers sometimes included inaccurate or nonsensical sentences.

Without details of the quality of the data used to train ChatGPT and other large language models, it is hard to gauge the scale of such bias. Under-studied languages could end up being excluded from the chatbot revolution. Dialogue between developers and users of large language models must be transparent to ensure that this does not happen.

Mohamed L. Seghier Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates.
mseghier@gmail.com

Brazil: plan for zero vegetation loss in the Cerrado

Brazil's new government aims to achieve zero deforestation in the Amazon rainforest. This welcome initiative should be extended to the Cerrado, the world's most biodiverse woodland savannah.

The Cerrado covers 2 million square kilometres and is home to more than 5% of all Earth's known plant and vertebrate species. But agricultural monocultures have expanded steadily across the region, which is now a major producer of commodities, with the loss of almost half its area of natural vegetation. The extent of deforestation exceeds that in the Amazon and in the Atlantic Forest. Pesticides, excessive water use and dam construction have damaged the Cerrado's freshwater systems. The rapid and aggressive occupation of the region's lands has led to conflicts with Indigenous people and other traditional populations.

The Cerrado's remaining protections for natural vegetation depend on the consolidation of an existing regional conservation network that connects public and private protected areas, Indigenous lands and portions of private lands that require special protection under Brazilian legislation. This will not harm Brazil's economy, because modern technologies and improved land use can boost the region's agricultural output without threatening its biodiversity.

Ricardo B. Machado, **Ludmilla M. S. Aguiar** University of Brasilia, Brasilia, Brazil.
rbmac@unb.br

José Maria C. Silva University of Miami, Coral Gables, Florida, USA.

PhD training: exposing obstacles to reform

Training for a PhD must be reformed if it is to meet society's expectations (see *Nature* **613**, 414; 2023). At Lehigh University, we investigated the feasibility of implementing a model solution known as the Pasteur Partnership PhD, which hinges on use-inspired research and professional development in partnership with industry (see go.nature.com/3jo7pfz).

More than 70% of the students admitted to our science, technology, engineering and mathematics PhD programmes in the past year were interested in this option but only 3% enrolled, because industry partnerships are hard to organize. Governments should help to meet this training challenge, which is driven by societal needs.

Senior faculty members often see no need to change the system. Younger researchers are willing to adapt but are constrained by tenure, promotion and funding expectations that reward conventional research output. Academic performance criteria need to be revised to allow training time and to incorporate innovative research.

University leaders are too often preoccupied with intellectual property and an open research ethos. And short-term profitability by publicly traded companies takes priority over industrial research and its associated financial commitments. To 'turn the PhD tanker', these misalignments must be addressed.

Himanshu Jain, **Nathan Urban** Lehigh University, Bethlehem, Pennsylvania, USA.
h.jain@lehigh.edu

Gary S. Calabrese Wilmington, North Carolina, USA.

HOW TO SUBMIT CORRESPONDENCE

Please consult the full author guidelines and policies at go.nature.com/cmchno before e-mailing your submission to correspondence@nature.com. It should not be sent as an attachment.

The following cannot be considered: technical comments on peer-reviewed research papers; responses to articles published in journals other than *Nature*; contributions that present primary research data; and submissions that do not comply with the section's strict length and style requirements. Submissions must not be under consideration elsewhere.

Please make clear in your proposed text its pertinence to *Nature*'s global readership. Ensure that you include a link to any article under discussion and references or links for fact-checking of all statements, in addition to the limited number of references necessary for publication.

© 2023 Springer Nature Limited. All rights reserved.

Work

Your story

Send your careers story to: naturecareerseditor@nature.com

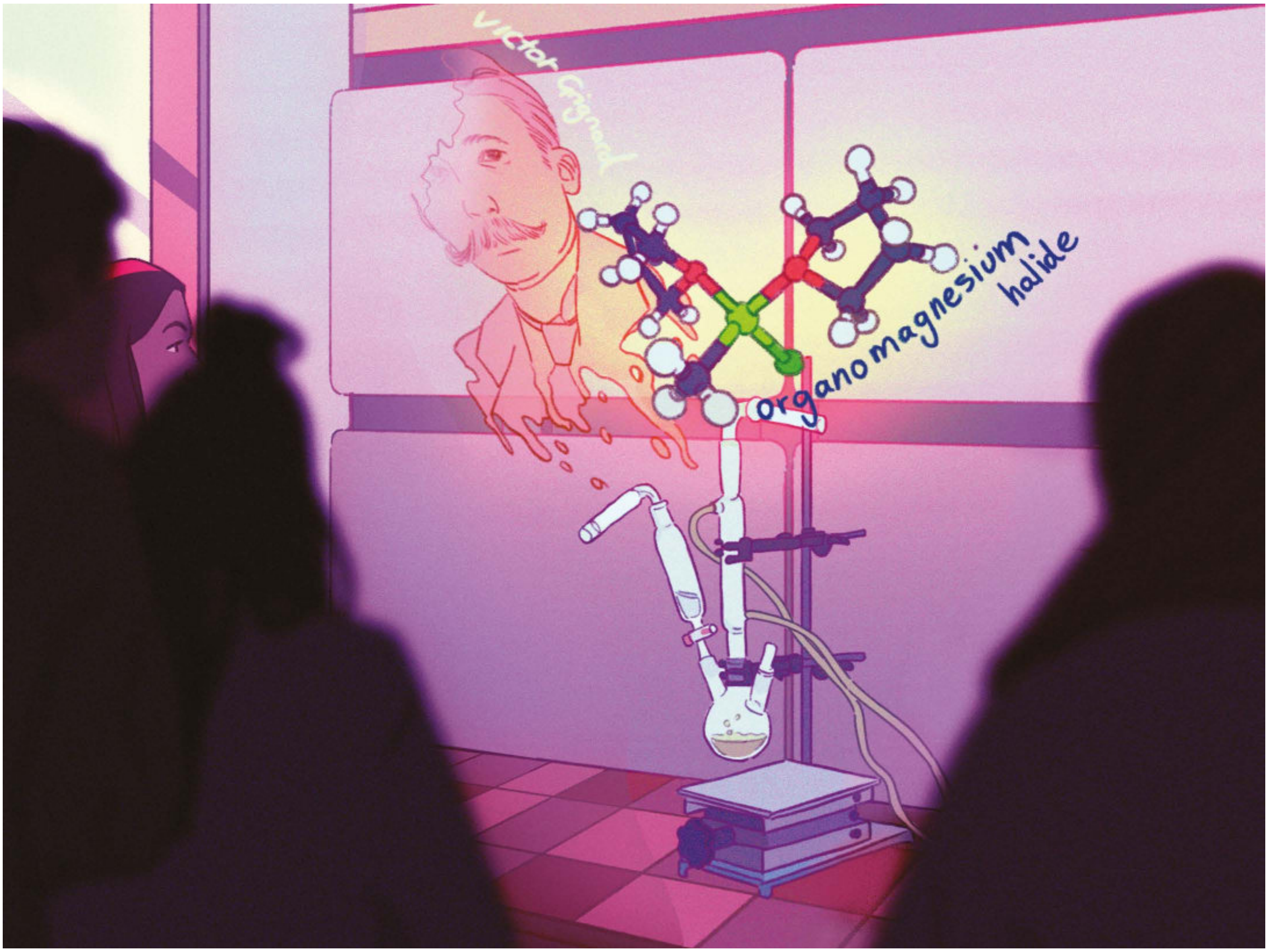


ILLUSTRATION BY BEX GLENDINING

Replacing historical chemical names, such as the Grignard reagent, with descriptive ones helps to decolonize chemistry courses.

CHEMISTRY REACTS TO CHANGE

Textbooks imply that chemistry is done only by white men. Curricula must change to better reflect reality. **By Katharine Sanderson**

Flick through a chemistry textbook at school or university, and you'd be forgiven for thinking that chemistry – touted as the central science that connects all other physical-science fields – is the domain of almost exclusively white men. Yet chemistry has origins going back millennia – such as in the papyrus-making procedures of ancient Egyptians, Chinese medicine and distillation methods developed in early Arab cultures. Contributions to modern chemistry are truly global, meaning that low- and middle-income countries can take advantage of chemical advances in areas such as agriculture and medicine as they seek to develop their economies.

To better reflect this breadth, chemistry departments across the globe need to revamp curricula, say those who are pushing for change. It must and will happen, says Avtar Matharu, director of the master's course in green chemistry at the University of York, UK. Matharu, who is Sikh, was born in Nairobi and moved to the United Kingdom as a child.

© 2023 Springer Nature Limited. All rights reserved.

He is also vice-chair of the university's staff race equality forum. "Universities are institutionally racist," he says, but adds that calls for change are too strong for institutions to ignore. "Every university has to respond to the issue," Matharu says.

At the University of York, the chemistry department is an early adopter of a decolonizing agenda and is already known for its progressive stance on inclusion and diversity. This includes LGBT+ friendly policies that take into account the needs of people from sexual and gender minorities; diversity and inclusion training for all undergraduate and postgraduate students; and unconscious-bias training for all recruitment panels. In 2007, it was the first department to achieve a gold Athena Swan award, the highest honour of the UK-wide programme, which recognizes significant gender-equity efforts in the sciences.

In 2019, York students from a variety of disciplines started to urge senior faculty members to decolonize their curricula. This means ensuring that there is a diverse community in departments, that course materials encompass sources from all parts of society and the world, and that courses reflect the lived experience of the diverse student body. In the academic year 2012–13, 21% of York students identified as Black and minority ethnic, rising to 38% in 2022–23.

Within a few months, students on chemistry courses saw that progress was being made in the humanities, but less so in science. So they called on the head of chemistry, Caroline Dessent, to act. She says that the chemistry students' call for action came when the department had already begun to think about how it could address racism and race-related discrimination at all levels. "There was also already quite a desire to do something to improve the culture in the department" for the minority ethnic students and staff, says Dessent, who is white.

In the city of York, 93% of residents are white, according to 2021 census data. "The environment is not very diversified," says Kelechi Uleanya, a postdoctoral research associate in Dessent's research group. Uleanya hails from Nigeria, where she completed her undergraduate and master's studies in chemistry. When she was working as a graduate teaching assistant (GTA) in the undergraduate laboratories at York, she noticed that other GTAs could have benefited from some training in how cultures take different approaches to practical work.

"Some [GTAs] might not have had serious contact with people of different cultures" and therefore lack context about why students might engage in problem sets or labs in different ways, she says. "So, for some of them it's a shock meeting [students from different cultures] in the labs and their approach to things is totally different because they are coming from another setting."



EDINBURGH NAPIER UNIV.

Nazira Karodia calls racism's negative effects on higher education "pervasive" for students.

Uleanya joined a steering group, made up of staff and students, that Dessent started in 2020 in response to students' concerns.

Dessent then wrote to all staff members setting out the decolonizing aims. "It is extremely important when you do equality, diversity and inclusion work that you explain why you're doing it, because otherwise you're always at risk of people trying to criticize it," she says. "Decolonizing is one of those things that currently does attract a certain amount of negative publicity. Maybe because people just don't understand what they're trying to achieve through it. It's really something very pragmatic."

Distilling lessons

In her newsletter to staff, Dessent laid out why the department was embarking on this course of action. Nationally, there is a 16.1% gap between the number of highest-honours degrees awarded to white UK students compared with Black and minority ethnic UK students, according to 2011 data from the UK National Union of Students (NUS), and this attainment gap is also present at York. Dessent noted that, in the NUS survey, 42% of Black students say that the curriculum does not reflect issues of diversity, equality and discrimination. Ten per cent of transgender students and nearly 25% of women never feel comfortable to speak up in class. "Our undergraduate courses should not disadvantage any student because of their background or characteristics. All of our students should have equal opportunities to thrive in our department," she wrote.

Steps towards achieving this include changing the taught material to be more reflective of contributions from scientists other than those who are white and male, but it also means working to change the culture in the department, she says.

Dessent asked all of her department's faculty members to scrutinize the courses they were currently teaching and find places to introduce some context that would help to diversify the curriculum. For instance, in a course on colloid chemistry that discusses the applications of micelles and gels in cleaning and lubrication, a lecturer incorporated how ancient Babylonians produced one of the first known lubricating gels used as chariot axle grease. In a medicinal chemistry course that covers a global history of medicine, instructors included examples from China, India, the Arab world and Indigenous Australian cultures.

Faculty members were also encouraged to include the work of individual non-Western chemists. For example, the medicinal chemistry course highlights the Chinese scientist Tu Youyou, who was the first scientist working in China to receive the Nobel Prize in Physiology or Medicine. She developed the antimalarial drug artemisinin by using her knowledge of traditional Chinese medicine and combining it with modern drug-discovery techniques. "Youyou isolated the active ingredient and did all the medicinal development from that traditional herbal drug to turn it into a pharmaceutical drug," says David Smith, who is white British and a prominent advocate for equity, diversity and inclusion in York's chemistry department. He points out that, ironically, the production of that drug "is now controlled by Western companies, and still isn't available in sub-Saharan Africa", where it is most needed to combat malaria. Teaching this context as part of a pure chemistry course adds a globalized aspect to the curriculum that shows students how chemistry intersects with traditional

© 2023 Springer Nature Limited. All rights reserved.

knowledge, culture and politics, he says.

These examples, along with many others, have been collated into an evolving document by Uleanya and her colleagues (O. K. Uleanya *et al.* Preprint at ChemRxiv https://doi.org/jzfb; 2022).

Uleanya has been compiling and annotating the resource since October 2021 with funding from the Royal Society of Chemistry in London. "It's a detailed resource that anybody can go and pull examples from" for teaching high-school or university-level courses, Uleanya explains. When she joined York in August 2016, she was pleased that the department had a reputation for being engaged in equity, diversity and inclusion work, but she still wanted to find more role models. She began to take a deeper look at the literature and found that there were chemists making important discoveries from other countries and other cultures. "People like me contributed a lot," she says.

And yet in chemistry, lots of the reactions that students learn are named after the often-white male chemist who is credited with discovering or developing them. "There is no need for the name of a reaction to reflect a person," says Smith. Instead, those reactions can be taught by their reactivity – what actually happens if that kind of reaction is performed on a certain kind of molecule. For example, the Grignard reaction, which is used in the formation of carbon–carbon bonds, is named after French chemist Victor Grignard. The reaction starts by making a Grignard reagent, a molecule containing a halogen, a magnesium atom and an organic group, which could more descriptively be called an organomagnesium halide, Smith says.

For many students, this is likely to make more sense – removing the need to simply memorize names and instead think more deeply about the chemistry that the reactions exemplify. "Very often, by decolonizing the name you can make things clearer," says Smith. But wholesale rewriting and reprinting of textbooks, as well as adopting a different name for a historically well-known procedure, is likely to be a slow process, he adds.

Smith emphasizes that the department isn't trying to dictate to lecturers what they must include in their teaching material. Instead, "it has been a bit more about role modelling the contexts of chemistry", he says. For instance, lecturers could use examples of chemistry that solves a problem in low-income countries, or of science done by researchers outside the countries that conventionally dominate the literature, he suggests.

Charting prejudices

Matharu isn't convinced that this kind of collation exercise is ambitious enough. "The easiest thing to do is find lots of examples in the literature for minority ethnic researchers," he says, but describes this approach as too simple. If a student had a follow-up question, and the teacher or lecturer had included the example as a box-ticking exercise, then the instructor would probably have no further context to add. Matharu has developed his own teaching methods that lead students to realize for themselves how implicitly biased science and the scientific literature are. "I was very keen that we need to start the conversation," he says.

For example, Matharu introduced a section in written essays, in which students have to look up the country and institution of the corresponding authors on all the papers they cite and construct a bar chart. Inevitably, these plots are dominated by US, EU and UK researchers. He first introduced this after marking a student project on coffee production, which predominantly happens in the global south. Matharu totted up the countries of the corresponding authors and realized that there were hardly any from the countries where coffee is actually produced. Now, for all the projects in his postgraduate course on green chemistry, the students must write a narrative about the trend they see in cited authorship.

"Universities are institutionally racist. Every university has to respond to the issue."



CHRISTINA SURDHAR

Avtar Matharu helps students identify biases.

"That's really eye-opening," says Matharu. "When they start to write their narrative, or even see the plot, the plot should already encourage them that their literature is biased." Matharu now gives points, which count towards a student's final degree mark, for completing this exercise. "It's not something that's too onerous," he says, "but it's something that starts to change their behaviour; it gets them to think."

One student, he says, wrote that the review made them think about the origin of the papers on a deeper level. "That's quite a profound statement," he says. "For that particular student, I changed their mindset and changed their way of thinking." He has dubbed the authorship charts 'Matharu plots'.

Arinzechukwu Ngoka, who is Black and studied chemistry as an undergraduate in his home country of Nigeria, came to York to pursue the green chemistry master's course. Ngoka says that the graphs are really easy to compile, but that the results are stark. "Matharu plots are an eye-opener to my open bias of Western literatures. I was used to the traditional and biased system thinking that the white scientists dominated the educational system," he says, adding that he now searches for papers from a more diverse range of authors than he did previously.

Matharu hopes that other institutions will take on his idea, and he has presented it to the Royal Society of Chemistry and to some major publishers, he says.

York doesn't stand alone in its decolonization efforts. In her previous roles as pro-vice chancellor with responsibility for regional engagement and as a science-education professor at the University of Wolverhampton, UK, Nazira Karodia led a range of decolonizing and diversity and inclusion programmes in science and technology. Now deputy vice-chancellor at Edinburgh Napier University, UK, Karodia grew up during the apartheid period of systemic, nationalized racial discrimination in South Africa. "I have had first-hand experience of racism in South African education and society and have seen the pervasive negative influence of racism in higher education in England – a racism that affects so many students' study, lives, expectations and futures," says Karodia.

Karodia says that maintaining a dialogue with stakeholders is key to implementing a successful decolonizing agenda. Students seem to accept change more readily, in her experience, but "staff need to be persuaded, sometimes, by rational argument and an appeal to the values of the liberal institutions. Some staff will never be convinced, and we have to live with that too", she says.

Uleanya has showcased her chemistry-decolonization resource at a number of conferences on equity, diversity and inclusion in science, and interest has been high, she says. A common concern that she hears is that decolonizing the curriculum means a loss of quality. Her advice is to start slowly, including a few new examples. "It has nothing to do with losing the chemistry we have," she says, but rather about including things to make chemistry more global. "Where everybody can feel a sense of belonging. And then to build better chemists."

Katharine Sanderson is a locum senior reporter at *Nature*.

© 2023 Springer Nature Limited. All rights reserved.

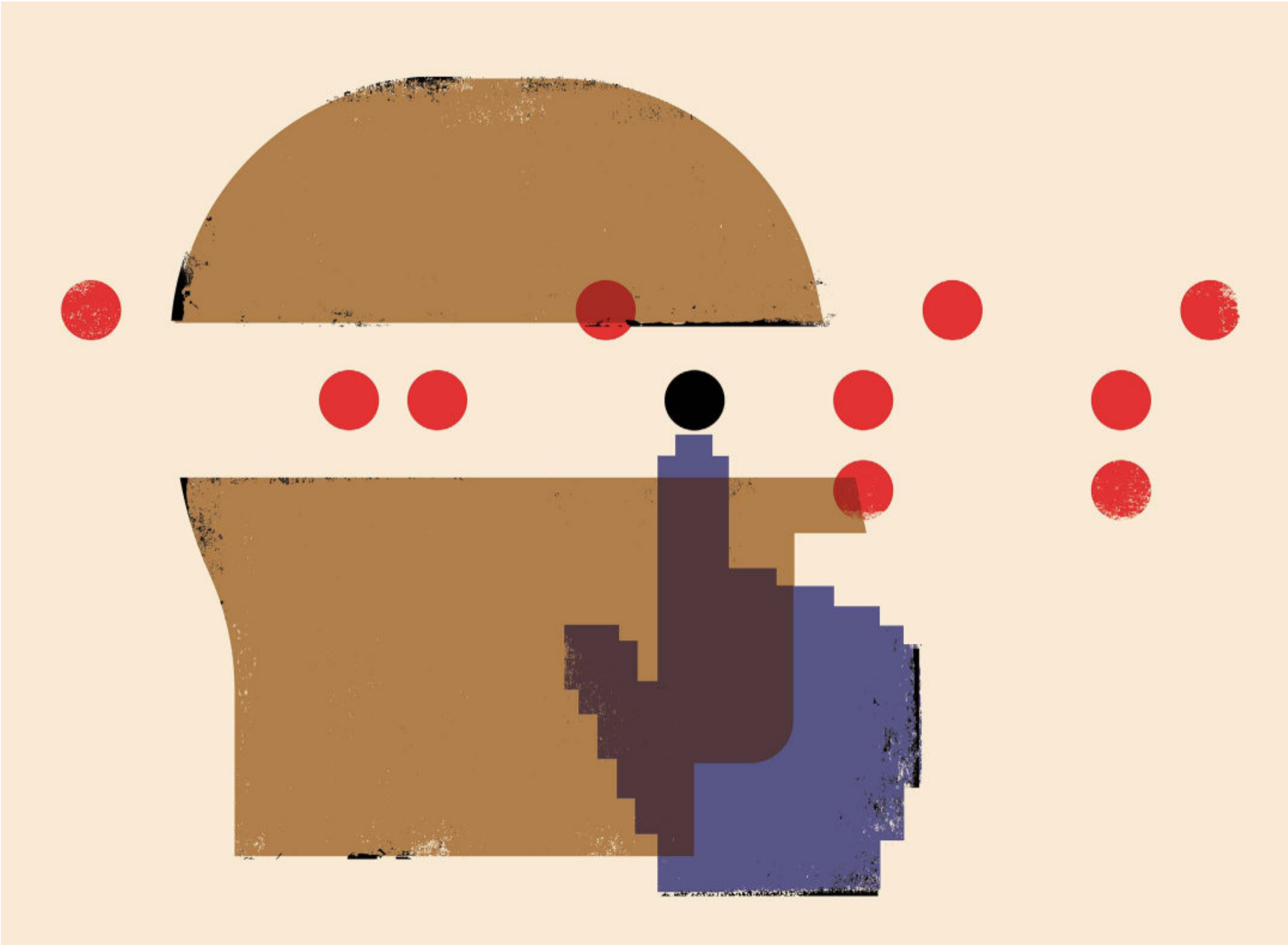


ILLUSTRATION BY THE PROJECT TWINS

TOOLS FOR READING PAPERS WITHOUT THE BENEFIT OF SIGHT

How innovative software and modes of presentation can help researchers with visual disabilities to access the literature. **By Alla Katsnelson**

Nearly 240 data scientists have proved themselves nimble enough at their work to be certified instructors for the 'tidyverse', a popular package for manipulating and visualizing data in the R programming language. JooYoung Seo is unique among them – the first blind instructor to gain certification.

Most tidyverse users present data in the form of charts and graphs. Seo, an information and learning scientist at the University of Illinois at Urbana–Champaign, who lost his sight at age ten owing to glaucoma, uses touch, sound and speech. He is one of a small but growing group of researchers working to make science more accessible to people with limited vision. "The overarching challenge is that content is visually designed," Seo says. "But visualization is only one of the representation methods. We can represent data in multimodal ways."

There are no good estimates of how many scientists with low vision are working today, but a 2020 study[1] found that fewer than 100 of 52,124 researchers applying for funding from the US National Institutes of Health (NIH) in 2018 self-identified as having a visual impairment. "It's a fraction of a fraction," says Bonnielin Swenor, an epidemiologist and director of the Disability Research Center at Johns Hopkins University in Baltimore, Maryland, who led that work. That's far short of the total number of scientists with vision disabilities, Swenor adds, because ableist hurdles prevent many scientists from applying for research grants. And it's an even smaller fraction of the number of individuals with visual disabilities overall: in 2017, some 7 million people in the United States (2.17% of the population) were living with 'uncorrectable' loss of visual acuity or with blindness[2]. "If the goal is parity with prevalence in the US – which I argue it should be – we aren't close," she says.

Avoiding workarounds

That's in part because, all too often, even the most basic research activities – accessing the literature, submitting papers, attending conferences and reviewing manuscripts – require time-intensive, personalized workarounds. "Blind scientists probably have a way that works for them," says Daniel Hajas, an

© 2023 Springer Nature Limited. All rights reserved.

innovation manager at the Global Disability Innovation Hub at University College London who works to accelerate assistive technologies to market. “But it’s not generalizable and it took them a long time to get there.”

Swenor, who has degenerative retinal disease, recalls that a mentor early in her career advised her that she would have to work four times as hard as sighted colleagues to succeed in science. “That should not be the path, for anyone,” she says. But awareness of the need for accessibility in science is growing, and new tools are on the way.

People whose visual acuity is too low for them to read magnified text generally use screen readers, which convert digital text into either synthesized speech or a tactile (Braille) display.

Screen readers navigate digital documents using tags that identify elements such as headers, figures, tables and footnotes. But the devices often stumble on PDF files, which have long been the default digital format for journal articles and other research materials, because they generally lack such tags. The two-column format widely used in journals can also confuse screen readers, which generally read from left to right across a page. PDFs can be made accessible, but it’s challenging to do and many publishers haven’t got there yet. “Improving PDF accessibility requires changes to culture, systems and processes that can be challenging for publishers to achieve,” says Jude Robinson, a global lead for Springer Nature Digital in London. “We are committed, however, to doing this.” (*Nature*’s news team is editorially independent of its publisher, Springer Nature.)

An online tool called SciA11y, developed by the Allen Institute for Artificial Intelligence (AI2) in Seattle, Washington, uses machine learning to extract the content and structure of a PDF (and other file formats, such as LaTeX) and re-render it in HTML, creating a table of contents containing links to tagged section headings that can be navigated with screen readers. AI2 has also built in functionality such as bidirectional navigation between in-line citations and their corresponding references in the bibliography, says Jonathan Bragg, who co-leads the project.

In a detailed survey[3], six scientists with vision loss described how they frequently found themselves unable to access or read PDFs. One respondent mentioned that they encountered problems two-thirds of the time, and that they use at least six different approaches to read papers. “It was eye-opening to see the range of tools that people use when reading — and the struggles they have when those papers have not been formulated appropriately,” Bragg says.

At the moment, SciA11y — ‘a11y’ is Internet shorthand for accessibility — is an online demo: researchers can upload PDFs to re-render them in HTML. But the team is still working on key functionality, Bragg says. For example, the software still makes mistakes, sometimes failing to pull out headings and leaving them as body text in the HTML. It also struggles with tables and images, which is a current focus of development.

In that respect, SciA11y is not unique: for visually impaired people, tables and images present challenges that are even harder to solve than text itself. Precious little scientific literature — whether in PDF, HTML or another format — includes textual descriptions of figures, called alt text, that would enable a blind or low-vision person to understand the images. What’s more, it’s not always clear what sort of explanation would be most relevant. “There are different schools of thought,” says Amy Bower, a blind physical oceanographer at the Woods Hole Oceanographic Institution in Massachusetts. “Should you just describe what’s there, or should you add the interpretation?”

> **“I can hear the trend of the data in sound, and feel the pattern.”**

Some researchers, including Bower, solve this issue by working with sighted interpreters. Hajas is collaborating with researchers at the Massachusetts Institute of Technology in Cambridge and others to design Olli, a screen-reader interface that allows users to navigate up and down different levels of description — from a single sentence explaining a figure’s main takeaway, to a more detailed characterization of the axes and legend, to the actual data values from which it is built. Olli currently supports basic chart types, such as bar charts and scatterplots, and Hajas says that the team is working on new features, such as heat maps.

Seo is also developing a data visualization tool, and he is drawing on more than just speech. The tool, called the multimodal access and interactive data representation system (MAIDR), encodes data as both sounds — termed sonification — and Braille, providing tactile analysis with the help of a refreshable Braille display. “I can hear the trend of the data in sound, and feel the pattern,” Seo says.

Experiencing data through touch can be especially powerful, says Mona Minkara, a computational chemist and bioengineer at Northeastern University in Boston, Massachusetts, who began to lose her vision as a young child. Minkara collaborated on a 2022 study[4] that described representing data as 3D-printed graphics called lithophanes. Produced from plastic thin enough for light to shine through, lithophanes can encode multiple forms of chemical and life-science data — for example, a scanning electron micrograph of a butterfly wing, the bands of an electrophoresis gel or the ultraviolet spectrum of a protein. The technology, she says, allows her and her sighted laboratory colleagues to engage with the same data at the same time.

Feeling a change in protein function through your fingers as a difference in thickness, for example, “adds a layer of knowledge”, she says, regardless of whether a researcher can perceive that change visually. “It’s going to embed into your imagination and be more a part of you, so you understand it on a deeper level.”

Sonification provides another way for scientists with visual limitations to ‘see’ data, for instance in astronomy. “We are used to thinking that astronomy is a visual science, but most of the data that we get are just numbers,” says Anita Zanella, an astronomer at the Italian National Institute for Astrophysics in Rome. “We translate them into images because that’s a way for us to make sense of them”, but ears are often better at detecting faint signals. By translating numerical values into sounds with certain parameters — for example, a star’s brightness might be encoded as pitch — researchers can home in on important changes[5].

Researchers in genomics and geology are also exploring sonification, although shared principles and standards that would allow scientists to compare data in auditory formats are still in development, Zanella says. Still, simple online tools, such as the Highcharts Sonification Studio, developed at the Georgia Institute of Technology in Atlanta, allow researchers in any field to upload data and explore ways to represent them aurally.

In addition to new tools, mindsets have advanced substantially over the past five years as researchers and developers increasingly consider accessibility, according to several researchers interviewed for this article. And in the United States, at least some federal funding bodies have made accessibility a priority. On 30 December, the NIH published recommendations (see go.nature.com/3gtmtk2) for breaking down ableism and creating equitable and inclusive research practices. (Swenor co-chaired the NIH’s Subgroup on Individuals with Disabilities, which developed the proposals.)

But for researchers with vision loss to truly succeed, tools will have to evolve beyond piecemeal efforts, Hajas says. “Unless we make the whole ecosystem accessible — how to find articles, how to read the text, how to access the diagrams and everything else — there will be missing pieces that essentially disable people from becoming good scientists,” he says.

Alla Katsnelson is a science writer based in Southampton, Massachusetts.

1. Swenor, B. K., Munoz, B. & Meeks, L. M. *PLoS ONE* **15**, e0228686 (2020).
2. Flaxman, A. D. *et al. JAMA Ophthalmol.* **139**, 717–723 (2021).
3. Wang, L. L. *et al.* Preprint at https://arxiv.org/abs/2105.00076 (2021).
4. Koone, J. C. *et al. Sci. Adv.* **8**, eabq2640 (2022).
5. Zanella, A. *et al. Nature Astron.* **6**, 1241–1248 (2022).

© 2023 Springer Nature Limited. All rights reserved.



Where I work
Samantha Cristoforetti

Photographed by ESA/NASA.

I'm an astronaut with the European Space Agency. Last year, I spent five months on the International Space Station (ISS) – from late April to mid-October – and spent the last month as station commander. Before returning to Earth, my team and I set aside time for a water 'play date'. Here, inside the ISS, I am demonstrating how water behaves in zero gravity.

You can use little tricks to make sure the water stays where you want it. Surface tension keeps the water bubble together, and you can move it by using a straw to pull on it gently, or by blowing on it. If the bubble is small enough, you can drink it. We recycle all the water onboard.

Weightlessness is not just exhilarating, it's also an opportunity to study fundamental physics. There's a lot of research on the space station on fluid dynamics. One study I was personally involved in looked at the sloshing behaviour of different kinds of fluid and mixes of fluids and gases in containers. The results are important for designing fuel tanks, especially for use in space, but also on Earth.

The photo was taken in the Japanese Experiment Module. This is the largest single ISS module and the least cramped, so we often use it for talking to the media or to schoolchildren. When we communicate with them, we use props, such as the balls behind me, which are models of planets and the Moon. The round thing behind me is the module's airlock. We use it for deploying satellites, as well as hardware such as scanners for science experiments.

This was my second time on the ISS. I adapt quickly to space – I really enjoy that feeling of weightlessness. It's a lot harder for me to come back to Earth.

I don't know when I'm going up there again – or if I will. We'll see how the US-led Artemis programme – which will return people to the Moon in the coming decade – evolves. Maybe there will be an opportunity for me.

Samantha Cristoforetti is a European Space Agency astronaut who lives in Cologne, Germany. **Interview by Linda Nordling.**

© 2023 Springer Nature Limited. All rights reserved.

News & views

Neuroscience

How an anxious heart talks to the brain

Yoni Couderc & Anna Beyeler

During periods of anxiety, the brain affects the heart, but does a racing heart also talk to the brain to cause anxiety-related behaviour? Use of a light-stimulated pacemaker in mice shows that it does, and pinpoints a brain region involved. **See p.292**

Have you ever been so anxious that you could feel your heart racing in your chest? This tachycardia is one of the main symptoms of anxiety[1], and can be so intense that the person experiencing it sometimes mistakes it for a heart attack. Experimental research has revealed numerous pathways that convey signals from the brain to the heart. But in both clinical psychiatry and fundamental neuroscience, the reverse – the effect of heart rate on emotions – has remained a debated question for almost a century[2]. Hsueh *et al.*[3] address this issue on page 292 , identifying a mechanism by which the brain detects heart rate, and showing how this, in turn, controls emotional behaviour.

Interoception is the continuous perception by the brain of internal signals in the body, including those from the respiratory, gastrointestinal and cardiac systems[4]. In people who have anxiety disorders, sensitivity to these internal signals – especially heart rate – is altered[1]. Studies in animals have highlighted the link between cardiac changes and emotional states[5,6], but whether an increased heart rate contributes directly to anxiety had remained unclear.

So far, only nonspecific electric shocks, vagal-nerve stimulation and pharmacological approaches – all of which involve major side effects – have been used to increase or decrease heart rate and to evaluate its effect on emotions. Researchers have lacked tools with the necessary temporal and spatial resolution to adequately investigate the effects of heart rate on anxiety.

The first of Hsueh and colleagues' breakthroughs was the development of such a tool: a non-invasive optical pacemaker. It was based on the systemic delivery into mice of a viral vector carrying a gene that encodes a light-sensitive protein, the opsin ChRmine. When illuminated with red light, positively charged ions flow through this protein, which depolarizes cells that express it. In the authors' experiments, the cells targeted were heart muscle cells, the depolarization of which triggers muscle contraction.

By mounting a red micro-light-emitting diode (micro-LED), blinking at a defined frequency, to a vest worn by mice, the authors could apply their 'optogenetic' strategy to control heart rate (Fig. 1). ChRmine has been used previously to precisely control specific neural circuits in deep regions of the brain without the need for intracranial surgery[7]. Hsueh and colleagues have extended the applications of this molecular tool by using it to control the activity of (to pace) an entire organ and to determine any heart-to-brain influences on anxiety.

The authors applied their approach to test whether an increase in heart rate to 900 beats per minute (36% higher than the baseline frequency) could alter levels of anxiety in freely behaving mice. Hsueh *et al.* used two methods to assess anxiety – placing the animals in a maze or in an open field, both of which include safe and exposed areas. The authors found that optically induced tachycardia led to a greater avoidance of exposed areas in both assays, reflecting an increase in anxiety-related behaviour. This is an unequivocal demonstration that, at least in mice, heart rate can affect anxiety, and can probably influence

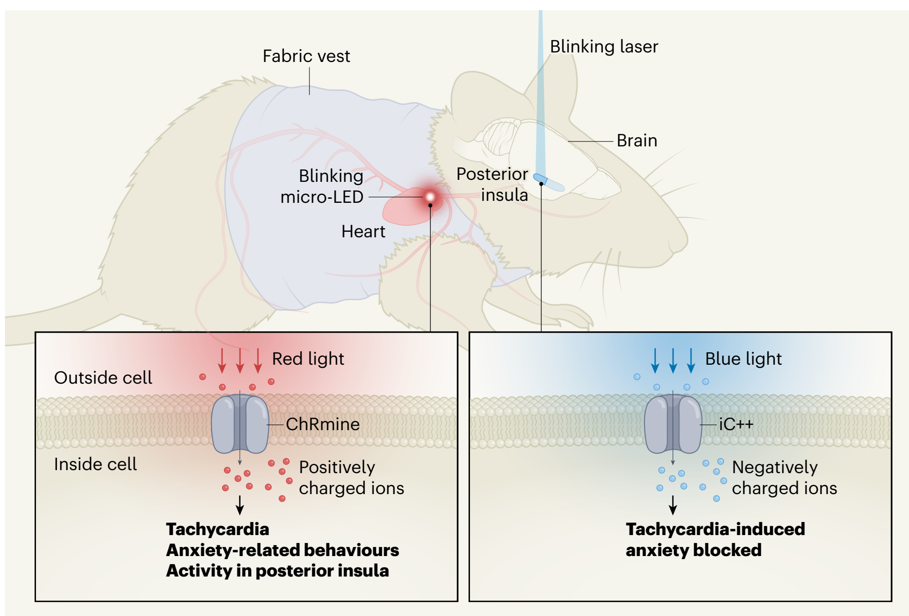


Figure 1 | Controlling anxiety with a non-invasive optical pacemaker. Hsueh *et al.*[3] studied the effects of a racing heart on anxiety-related behaviour in mice. They injected the animals with a heart-selective viral vector encoding a red-light-sensitive opsin protein called ChRmine. They dressed the mice in a vest containing a red micro-light-emitting diode (micro-LED), blinking at 900 beats per minute, that was fastened to the animals' chests over their hearts. The red light activated ChRmine in heart muscle cells, opening up the protein to let positively charged ions flow through it. This caused depolarization of the cells and an increase in the animals' heart rate (tachycardia). This, in turn, led to an increase in anxiety-related behaviours, and activated the posterior insula (among other brain regions). Hsueh *et al.* then inhibited the posterior insula using a blue-light-sensitive opsin, iC++, through which negatively charged ions flow. They activated iC++ using a blue laser, while optically increasing heart rate. This prevented tachycardia-induced anxiety.

© 2023 Springer Nature Limited. All rights reserved.

other emotional behaviours, too.

To investigate the neurobiology that underlies this tachycardia-induced anxiety, Hsueh *et al.* performed a screen of brain activity after 15 minutes of optically induced tachycardia. Whole-brain mapping of neurons revealed changes in gene expression in response to tachycardia. The authors found that neurons in two regions – the posterior insular cortex (posterior insula) and the brainstem – were strongly activated. Electrophysiological recordings in live mice also showed an increase in the firing rate of neurons in the posterior insula during optically induced tachycardia.

The insular cortex is involved in both interoceptive processing and anxiety-related behaviours[8–10]. By this stage, the authors had discovered a correlative increase in the activity of the posterior insula after an increase in heart rate. But any involvement of this region in the tachycardia-induced anxiety remained to be determined. To investigate this, Hsueh *et al.* optogenetically inhibited posterior insula neurons using a different opsin – the blue-light-sensitive protein iC++.

In so doing, they made a second discovery: inhibition of the posterior insula during optical pacing reduced the anxiety behaviours induced by tachycardia. This indicates that the posterior insula relays information about heart rate to affect anxiety. The attenuation was specific to the posterior insula, and was not observed with optogenetic inhibition of a different region, the medial prefrontal cortex.

Overall, then, Hsueh *et al.* have found that increases in heart rate promote anxiety-related behaviours in mice, and that this is mediated through the activation of specific brain structures that include the posterior insula. The authors' comprehensive study raises new questions, and opens up areas for research. For example, the neural circuits and mechanisms that allow the posterior insula to be activated by tachycardia – as well as the circuits that induce anxiety behaviours – have yet to be identified.

Another unexplored aspect is the long-term effect of days (or weeks) of optically induced tachycardia – a question with substantial clinical implications. This study lays the foundation for testing whether chronic increases in heart rate induce long-term changes in the brain, which could underlie harmful levels of anxiety. Testing this hypothesis would raise technical challenges, because the micro-LED vest used here is not suitable for such long periods of stimulation.

Finally, from a translational and therapeutic perspective, it might be possible to design experiments to slightly decrease heart rate. Would this change reduce anxiety-related behaviours? Hsueh and colleagues' work has provided the means to investigate this prospect.

Yoni Couderc and **Anna Beyeler** are at Bordeaux University, Neurocentre Magendie, INSERM 1215, 33000 Bordeaux, France.
e-mail: anna.beyeler@inserm.fr

1. Paulus, M. P. & Stein, M. B. *Brain Struct. Funct.* **214**, 451–463 (2010).
2. Cannon, W. B. *Am. J. Psychol.* **39**, 106–124 (1927).
3. Hsueh, B. *et al. Nature* **615**, 292–299 (2023).
4. Azzalini, D., Rebollo, I. & Tallon-Baudry, C. *Trends Cogn. Sci.* **23**, 488–509 (2019).
5. Livneh, Y. & Andermann, M. L. *Neuron* **109**, 3576–3593 (2021).
6. Klein, A. S., Dolensek, N., Weiand, C. & Gogolla, N. *Science* **374**, 1010–1015 (2021).
7. Chen, R. *et al. Nature Biotechnol.* **39**, 161–164 (2021).
8. Etkin, A. & Wager, T. D. *Am. J. Psychiatry* **164**, 1476–1488 (2007).
9. Gehrlach, D. A. *et al. Nature Neurosci.* **22**, 1424–1437 (2019).
10. Nicolas, C. *et al.* Preprint at Research Square https://doi.org/10.21203/rs.3.rs-964107/v1 (2021).

The authors declare no competing interests.
This article was published online on 1 March 2023.

Structural biology

MicroRNA uses a gym to get fit for Dicer enzyme

Gunter Meister

The enzyme Dicer cleaves a type of RNA called a pre-microRNA to make the mature functional RNA. Structural evidence now sheds light on the catalytic mechanism involved and the role of a newly found RNA sequence termed GYM. **See p.323 & p.331**

Gene expression can be silenced through targeted pathways that depend on small RNAs. On pages 323 and 331, Lee *et al.*[1,2] provide insights into a key step in the maturation of such RNAs that is mediated by the enzyme Dicer.

RNA-dependent gene-silencing pathways are found in almost all eukaryotes (organisms whose cells contain a nucleus). Many of these systems are fuelled by immature versions of RNA that are often in the form of double-stranded RNA (dsRNA). Such dsRNA serves as the origin of small regulatory RNAs that include microRNAs (miRNAs) and short-interfering RNAs (siRNAs). In both cases, the dsRNA precursors are cleaved by a particular class of enzyme that is characterized by having what is termed an RNase III domain[3].

The miRNAs are processed from stem-loop-structured precursors, also described as a hairpin, and they require the consecutive action of two RNase III enzymes. In animals, the enzyme Drosha conducts the first cut and Dicer the second. Both enzymes define the ends of a short double-stranded miRNA intermediate from which one miRNA strand is selected and incorporated into the protein complex RISC, in which the RNA directly binds to a member of the Argonaute protein family.

The siRNAs are typically processed from the ends of long dsRNAs by only one enzyme, Dicer[4]. This scenario, however, requires that Dicer moves along the dsRNA. Therefore, Dicer enzymes can generally be divided into what are called non-processive and processive enzymes, depending on whether they generate more than one small RNA from a dsRNA

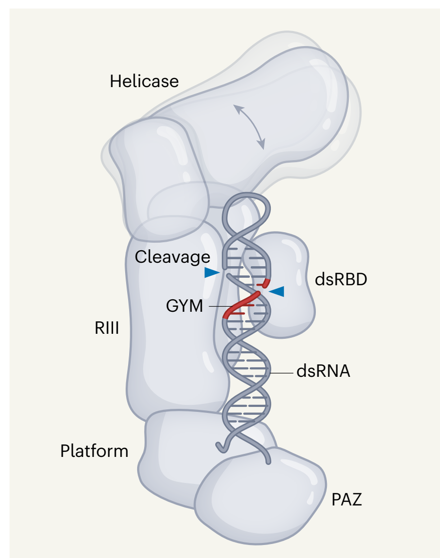


Figure 1 | Structural clues to how human Dicer enzyme functions. Some small RNAs, such as microRNAs, that function in gene silencing undergo a maturation step in which a double-stranded RNA (dsRNA) is cleaved by Dicer. Lee *et al.*[1] reveal that an RNA sequence that the authors term GYM has a role in facilitating this process. The authors' second paper[2] presents structural data obtained using cryo-electron microscopy, which captured the enzyme at the stage associated with RNA cleavage. This structure reveals the orientation of the dsRNA with respect to various domains of Dicer, a subset of which are shown here: helicase domain, RNase III (RIII) domain, dsRNA-binding domain (dsRBD), platform domain and PAZ domain. The helicase domain was not clearly visible in the structure, which suggests that it is in a flexible conformation at this step. (Adapted from Fig. 5 of ref. 2.)

© 2023 Springer Nature Limited. All rights reserved.

molecule. Human Dicer (hDicer) is specialized for pre-miRNAs and is thus non-processive. Although structures of Dicer enzymes from various organisms are available, the new studies of hDicer clarify the catalytic cleavage step (also known as 'dicing'), as well as revealing evolutionarily conserved pre-miRNA features that are important for efficient RNA processing.

The multidomain protein hDicer (Fig. 1) includes two RNase III domains, a helicase domain, a dsRNA-binding domain (dsRBD) and a PAZ domain, which, together with the platform domain, accommodates the two ends of the pre-miRNA hairpin[4]. Dicer not only cleaves the pre-miRNA but also functions as a 'molecular ruler' by essentially measuring the distance between the end of the RNA stem and the catalytic centres of the enzyme to produce a dsRNA that is 20–23 nucleotides in length. The sequences of pre-miRNAs are highly diverse, but besides the common RNA features of the hairpin structure, a two-nucleotide 3′ overhang on one side of the RNA (its 3′ end) and a phosphate group on the other side of the RNA (its 5′ end), no specific sequences or extra local structural motifs that might guide cleavage had been found previously.

To find such features, the first study by Lee *et al.*[1] used a pre-miRNA-processing approach in which RNAs were engineered to randomize the sequence of the upper stem of the pre-miRNA. Then the authors tested RNA-processing efficiency by sequencing the small RNA products produced after cleavage by Dicer. They thereby found an evolutionarily conserved sequence of nucleotides that they term GYM. This nucleotide motif, located around the cleavage site, consists of a paired guanine (G, one of the four main bases found in RNA), a paired pyrimidine (Y, which can be either one of the bases cytosine and uracil) and a mismatched (M) nucleotide pair. When this motif is engineered onto a random short hairpin RNA, Dicer processing of the RNA is markedly increased compared with the case for RNAs lacking the motif.

In their second study, Lee *et al.*[2] used hDicer and a pre-miRNA with a GYM motif as the material for generating structural data by means of cryo-electron microscopy. Strikingly, the authors found complexes with hDicer in the dicing state, which had not been seen so far.

Compared with structures in the pre-dicing state, in which the helicase domain binds to the pre-miRNA loop and keeps the pre-miRNA away from the catalytic centre, the helicase domain is not visible in the dicing-state structure. This observation indicates that the helicase domain becomes very flexible in the dicing state (Fig. 1). It might no longer bind to the pre-miRNA, enabling the pre-miRNA to move into a cleavage-competent position. Furthermore, in this position, the dsRBD of Dicer recognizes the GYM motif and stabilizes the interaction between the RNA and Dicer for a more efficient cleavage reaction. This molecular interaction explains the evolutionary conservation of the GYM motif.

Although several Dicer structures are available from different organisms and in different states[5–9], there are still many unknowns about this fascinating molecular machine. First, and most intriguingly, how is the cleavage product released in a post-dicing state? A RISC-loading complex has been postulated [10], consisting of Dicer, its partner protein TRBP and an Argonaute protein that takes over handling the miRNA from Dicer after cleavage. Major structural rearrangements and long-distance movements of the cleavage product are obviously required for this step. The helicase domain might become more static in such a complex, and TRBP might also contribute to structural rearrangements. Another question is how a correctly loaded RISC might be released from this complex.

Furthermore, we have not even begun to understand the role of alterations (termed post-translational modifications) of Dicer, such as the addition of phosphate groups (phosphorylation), in this process. Many such modifications have been reported, but if and how they contribute to the dicing cycle, and particularly to the structural flexibility of Dicer, is unknown.

Notably, both studies highlight cancer-associated mutations that correspond to prominent structural positions in Dicer. For example, the sequence encoding a pocket in the platform domain that accommodates the 5′ end of the pre-miRNA is often mutated in cancer. In addition, cancer-associated mutations in the dsRBD lead to reduced binding of Dicer to the GYM motif.

It is tempting to speculate that many more such mutations exist, which might not only affect cleavage by Dicer but potentially also later, post-dicing steps. What are the consequences of such mutations for cell growth and cancer development? Is there solely a reduction of global miRNA levels owing to impaired processing, or could there even be pre-miRNA-specific processing effects? Future mechanistic and structural work will provide further functional insights into the fundamental cellular process of miRNA formation and its link to disease.

Gunter Meister is at the Regensburg Center for Biochemistry, University of Regensburg, 93053 Regensburg, Germany.
e-mail: gunter.meister@ur.de

1. Lee, Y.-Y., Kim, H. & Kim, V. N. *Nature* **615**, 323–330 (2023).
2. Lee, Y.-Y., Lee, H., Kim, H., Kim, V. N. & Roh, S.-H. *Nature* **615**, 331–338 (2023).
3. Treiber, T., Treiber, N. & Meister, G. *Nature Rev. Mol. Cell Biol.* **20**, 5–20 (2019).
4. Torrez, R. M., Ohi, M. D. & Garner, A. L. *Biochemistry* **62**, 1–16 (2023).
5. Wang, Q. *et al. Science* **374**, 1152–1157 (2021).
6. Yamaguchi, S. *et al. Nature* **607**, 393–398 (2022).
7. Su, S. *et al. Nature* **607**, 399–406 (2022).
8. Jouravleva, K. *et al. Mol. Cell* **82**, 4049–4063 (2022).
9. Zapletal, D. *et al. Mol. Cell* **82**, 4064–4079 (2022).
10. Kobayashi, H. & Tomari, Y. *Biochim. Biophys. Acta* **1859**, 71–81 (2016).

The author declares no competing interests.
This article was published online on 22 February 2023.

Atmospheric chemistry

How wildfires deplete ozone in the stratosphere

V. Faye McNeill & Joel A. Thornton

Unexpected smoke-particle chemistry is shown to be the link between intense wildfires and stratospheric ozone loss. As the climate changes, more-frequent and more-intense fires might delay the recovery of the stratospheric ozone layer. **See p.259**

The devastating Australian bushfires of 2019–20 sent massive plumes of smoke high into the atmosphere, where it was transported around the world, affecting air quality as far away as South America[1]. Satellite data showed that this smoke also caused changes in the composition of the upper atmosphere, including a decline in stratospheric levels of ozone[2,3] — a gas that forms a protective layer around Earth, shielding terrestrial life from damaging short-wave ultraviolet radiation. But the mechanism by which wildfire smoke might enhance ozone depletion has remained uncertain. On page 259, Solomon *et al.*[4] make the case that particulate matter from wildfire smoke contributes to the destruction of stratospheric ozone, and suggest how it does this.

In 1974, modelling first pointed to increasing levels of chlorofluorocarbons (CFCs, which were often used in aerosol spray cans, foams and cooling devices) as a source of chlorine radicals in the stratosphere that could

© 2023 Springer Nature Limited. All rights reserved.

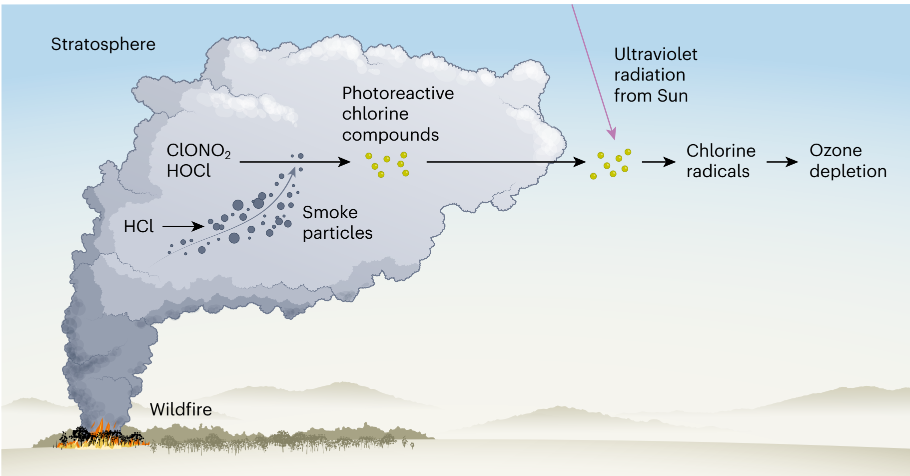


Figure 1 | A mechanism for wildfire-driven ozone depletion. Smoke particles from intense wildfires can enter the stratosphere. Solomon *et al.*[4] report that smoke particles in the stratosphere can absorb hydrogen chloride (HCl) gas, derived mostly from the breakdown of anthropogenic chlorofluorocarbon compounds (not shown). This allows the particles to catalyse the conversion of other chlorine-containing gases, such as chlorine nitrate ($ClONO_2$) and hypochlorous acid (HOCl), to 'photoreactive' chlorine compounds. When irradiated by ultraviolet light from the Sun, the photoreactive compounds produce chlorine radicals that catalytically destroy ozone, depleting its concentration in the stratosphere.

catalytically destroy ozone[5]. The dramatic discovery in the mid-1980s of the seasonal 'hole' in the ozone layer over the Southern Hemisphere polar region led to the 1987 Montreal Protocol on Substances that Deplete the Ozone Layer, the phasing out of CFCs worldwide and the expected gradual recovery of the stratospheric ozone layer — which is still not complete, owing to the long residence time of CFCs in Earth's atmosphere. When the ozone hole was first observed, the challenge for atmospheric chemists was to explain its unique seasonal and spatial characteristics. At the time, observations clearly linked the hole's appearance to anthropogenic chlorine, but there was no chemical explanation for its intensity and timing.

A major breakthrough came from theoretical and laboratory studies that pinpointed reactions occurring on the surfaces of stratospheric ice particles[6,7], which make up the iridescent 'mother-of-pearl' clouds seen in polar regions. Throughout the dark polar winter, those surfaces efficiently catalyse reactions of hydrogen chloride (HCl, mostly derived from the breakdown of CFCs) with other relatively inert chlorine-containing gases, including chlorine nitrate ($ClONO_2$) and hypochlorous acid (HOCl), to release 'photoactive' forms of chlorine that become reactive when irradiated by light. The return of sunlight in the polar spring transforms those forms of chlorine into a burst of highly reactive chlorine radicals, which catalyse processes that cause ozone loss. The ice particles further contribute to ozone depletion by removing certain nitrogen oxides (N_2O_5 and HNO_3) from the atmosphere; these compounds act as reservoirs for nitrogen oxide radicals, which, if present, limit the efficiency of chlorine-catalysed ozone destruction[6,7].

The properties of ice make it particularly well suited to catalysing chlorine activation and to removing nitrogen oxides at the low temperatures of the stratosphere, but there are clues that other particle types might also have a role. The powerful eruption of Mount Pinatubo in the Philippines in 1991 injected large amounts of sulfur dioxide gas into the stratosphere. After being transported long distances and oxidized, this gas became a layer of reflective sulfuric acid particles over much of the globe. In addition to reducing global average surface temperatures by 0.5 °C for more than a year, these particles also catalysed ozone loss in the stratosphere — not just over polar regions, but throughout the mid-latitudes[8].

Fast-forward 30 years to the 2019–20 Australian bushfires, which were another source of stratospheric particles, but with a very different chemical profile. The heat from large fires can lead to efficient lofting of smoke and its gas and particle constituents to high altitudes, where they enter the stratosphere. Solomon *et al.* now report that particulate matter from wildfire smoke contributes to the destruction of stratospheric ozone by efficiently absorbing HCl gas, thereby enhancing chemical activation of ozone-destroying chlorine radicals (Fig. 1). The authors also provide a glimpse of the properties of wildfire smoke particles in the stratosphere, which are challenging to observe directly or to recreate in the laboratory owing to their compositional complexity and the extreme conditions of the stratosphere.

Particles in wildfire smoke consist mostly of soot and organic carbon material, and are not immediately obvious candidates for enhancing chlorine activation and nitrogen oxide sequestration. Under dry conditions, such particles are generally less reactive than in ice or aqueous solutions[9]. Moreover, particles that are made up mostly of oxygenated organic compounds (such as those in smoke), and that might otherwise take up liquid water to better facilitate chlorine activation, can be glass-like at the cold temperatures and dry conditions of the lower stratosphere[10]. Reactant gases would not dissolve and diffuse into glassy particles as well as they would into liquid droplets. However, efficient HCl ionization and chlorine-activation reactions can occur on frozen polar stratospheric ice-particle surfaces as a result of the formation of disordered surface regions under stratospheric conditions[11], and this could also be true for glassy particles.

The first clue that smoke particles in the stratosphere might not be in a dry, glassy state came from an analysis of data acquired by the Atmospheric Chemistry Experiment Fourier Transform Spectrometer, a project in which the atmosphere is monitored by a spectrometer on a space satellite. This suggested that the Australian wildfire particles in the stratosphere contain both oxygenated organic material and adsorbed water[2].

Solomon *et al.* now build on this observation, using thermodynamic arguments to assert that complex mixtures of oxygenated organic material are likely to be liquid-like, even at low stratospheric temperatures. They also draw on an intriguing, albeit limited, set of laboratory studies of HCl solubility in mixtures of organic acids and water. These show that the potential for such mixtures to absorb HCl is even higher than that for sulfuric acid (the more common component of stratospheric particles). In other words, smoke particles have a higher potential to promote ozone depletion than was previously thought. On this basis, Solomon and colleagues are able to reproduce the ozone depletion observed during the Australian wildfires, using numerical modelling that combines their proposed process with the known chemistry of chlorine activation and chlorine-catalysed ozone destruction.

The study makes it clear that chlorine activation by smoke particles is the missing piece of the wildfire ozone-depletion puzzle. However, the findings raise questions about the physical state and chemistry of smoke particles in the stratosphere, questions that should now be further explored with field observations and laboratory studies under realistic conditions. Are these particles liquid-like, glassy or in some other state? And is a liquid-like state necessary to catalyse chlorine activation efficiently?

The findings also raise further questions about proposals to intentionally inject particles into the stratosphere to reflect sunlight and cool the planet — a climate-engineering

© 2023 Springer Nature Limited. All rights reserved.

approach called solar-radiation management. It has been proposed that, by using particles other than sulfuric acid, negative side effects such as ozone depletion and stratospheric warming could be bypassed or reduced[12]. Solomon and colleagues' study underscores the possibility that various particle types, fresh or after chemical transformation in the stratosphere, might still promote ozone depletion for a few decades as stratospheric chlorine levels slowly decline. Any candidate material for stratospheric injection must, at a minimum, be studied carefully in the laboratory, under simulated conditions of the stratosphere.

The Australian bushfire season of 2019–20 was extreme, but large wildfires and the resulting pyrocumulonimbus clouds are likely to become increasingly frequent and intense as the climate changes. The resulting injections of smoke into the upper atmosphere might therefore slow or temporarily disrupt the recovery of the stratospheric ozone layer. In this context, Solomon and colleagues' findings emphasize the need for atmospheric chemists to better understand the properties and reactivity of common, but complex, atmospheric particle types, such as those produced from biomass burning, in the cold and dry upper atmosphere.

V. Faye McNeill is in the Department of Chemical Engineering and the Department of Earth and Environmental Sciences, Columbia University, New York, New York 10027, USA. **Joel A. Thornton** is in the Department of Atmospheric Sciences, University of Washington, Seattle, Washington 98195, USA.
e-mails: vfm2103@columbia.edu ; thornton@atmos.uw.edu

1. Li, M., Shen, F. & Sun, X. *Sci. Rep.* **11**, 12288 (2021).
2. Bernath, P., Boone, C. & Crouse, J. *Science* **375**, 1292–1295 (2022).
3. Solomon, S. *et al. Proc. Natl Acad. Sci. USA* **119**, e2117325119 (2022).
4. Solomon, S. *et al. Nature* **615**, 259–264 (2023).
5. Molina, M. J. & Rowland, F. S. *Nature* **249**, 810–812 (1974).
6. Solomon, S., Garcia, R. R., Rowland, F. S. & Wuebbles, D. J. *Nature* **321**, 755–758 (1986).
7. Molina, M. J., Tso, T.-L., Molina, L. T. & Wang, F. C.-Y. *Science* **238**, 1253–1257 (1987).
8. Brasseur, G. & Granier, C. *Science* **257**, 1239–1242 (1992).
9. Abbatt, J. P. D., Lee, J. K. Y. & Thornton, J. A. *Chem. Soc. Rev.* **41**, 6555–6581 (2012).
10. McNeill, V. F., Loerting, T., Geiger, F. M., Trout, B. L. & Molina, M. J. *Proc. Natl Acad. Sci. USA* **103**, 9422–9427 (2006).
11. Zobrist, B., Marcolli, C., Pedernera, D. A. & Koop, T. *Atmos. Chem. Phys.* **8**, 5221–5244 (2008).
12. Keith, D. W., Weisenstein, D. K., Dykema, J. A. & Keutsch, F. N. *Proc. Natl Acad. Sci. USA* **113**, 14910–14914 (2016).

The authors declare no competing interests.

Condensed-matter physics

Hopes raised for ambient superconductors

ChangQing Jin & David Ceperley

A hydrogen-rich compound has taken the lead in the race for a material that can conduct electricity with zero resistance at room temperature and ambient pressure — the conditions required for many technological applications. **See p.244**

A superconductor is a metal that conducts electricity without resistance. In our increasingly electrified world, the implications of such a material are astounding — just imagine transmitting electrical power over thousands of kilometres with essentially no losses. But as promising as superconductivity might sound, this state has been achieved only at low temperatures[1] or very high pressures[2], both of which are unsuitable for many applications. For this reason, finding a material that can superconduct under ambient conditions has long been a focus of materials research. On page 244, Dasenbrock-Gammon *et al.*[3] report possible evidence for remarkable progress in this search.

When a material is in a superconducting state, its electrons form bound pairs known as Cooper pairs. These pairs avoid further collisions, and therefore encounter less resistance, than do single electrons moving alone. In some materials, crystal-lattice vibrations enable these Cooper pairs to form because the movement of positive ions in the lattice draws the electrons together. Materials containing hydrogen are particularly amenable to this pairing mechanism because hydrogen is the lightest chemical element, which means it has the highest vibration frequency. According to theory[4], this high frequency should increase the temperature at which a material can superconduct.

In 1968, physicist Neil Ashcroft predicted that pure hydrogen could superconduct at room temperature[5]. However, hydrogen becomes metallic only when H_2 molecules are dissociated at intense pressures of around 500 gigapascals (1 GPa is 10^9 Pa)[6]. This is about five million times that of atmospheric pressure, and extremely difficult to achieve with current experimental techniques. Ashcroft later suggested that hydrogen-rich compounds could become superconducting at lower pressures than can pure hydrogen, owing to the chemical compression induced by the other elements[7].

Experiments have since shown that several polyhydride compounds transition to a superconducting state at temperatures higher than 200 kelvin (some 93 K below room temperature). These compounds include sulfur hydride[2], rare-earth hydrides[8,9] and alkaline-earth hydrides[10,11]. But the pressures required are still very high — typically, hundreds of gigapascals.

Other materials exhibit a state of 'unconventional' superconductivity induced by lattice vibrations, as well as other physical processes that set these systems apart from their conventional counterparts[12,13]. A process called doping, in which impurities are deliberately added to a material, can induce, modify or enhance superconductivity in these materials — changing a non-superconducting state into a high-temperature superconducting state.

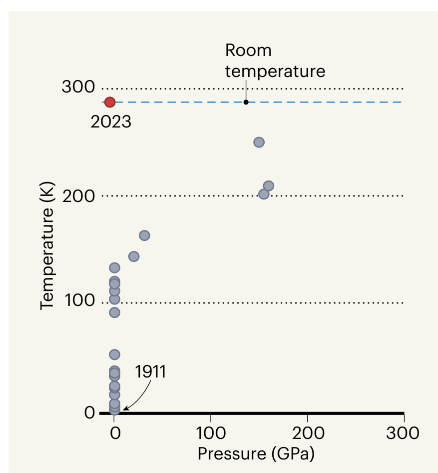


Figure 1 | The road to high-temperature superconductors. Researchers have been trying to achieve a superconducting (zero electrical resistance) state at ambient temperatures and pressures for more than a century. Several superconductors have been reported, albeit at very low temperatures. In 1968, it was predicted that pure hydrogen could superconduct at room temperature[5]. Dasenbrock-Gammon *et al.*[3] provide possible evidence that a hydride compound superconducts at 294 kelvin and 1 gigapascal (10^9 Pa), close to ambient conditions. (Adapted from Fig. 1 of ref. 20, with extra data from refs 10, 11.)

© 2023 Springer Nature Limited. All rights reserved.

Dasenbrock-Gammon *et al.* obtained evidence for near-ambient superconductivity by replacing some of the hydrogen in a lutetium hydride compound with nitrogen, which increased the number of charge carriers. The authors fabricated the compound at a pressure of 2 GPa. They then lowered the pressure to 1 GPa and measured a superconducting phase with a maximum transition temperature of 294 K. This is the highest temperature recorded at such low pressures (Fig. 1).

The sample underwent a striking visual transformation as it transitioned through different phases with changing pressure. When the non-superconducting metal was compressed, its colour changed from blue to pink, coincident with the onset of the superconducting regime at 171 K and 0.5 GPa. Then, when the authors compressed the sample to more than 1 GPa, the sample became non-superconducting again and its colour changed to a vivid red. It is intriguing that the superconducting state appeared only in the pink phase marked by intermediate pressures.

Dasenbrock-Gammon *et al.* made various measurements to show that this phase was indeed superconducting. They measured electrical resistance, and looked at how the voltage across their sample depended on the current through it. They also measured the temperature dependence of magnetic susceptibility (how easily a sample can be magnetized) and of a thermodynamic quantity called specific heat.

These measurements were all consistent and comprehensive. However, the authors' discovery will no doubt be controversial, because researchers from the same team previously retracted a report[14] of high-temperature superconductivity in carbonaceous sulfur hydride[15]. Independent measurements of the material, its properties and fabrication process will help to assuage any doubts about the results.

The authors suggest that hydrogen's high vibration frequency and the extra charge carriers from the nitrogen both contribute to the high-temperature superconducting state they measured. To determine whether this is indeed the case, more needs to be known about the composition and structure of the material. Dasenbrock-Gammon *et al.* used a technique called X-ray diffraction to show that the lutetium formed a closely packed crystal lattice, in an arrangement known as a face-centred-cubic structure, in some of the phases they detected. However, the locations of the hydrogen and nitrogen atoms could not be detected with X-rays. Future work is required to understand their distribution, which could be measured, for example, using neutron-diffraction methods.

The concentrations of hydrogen and nitrogen are also unknown. Dasenbrock-Gammon and colleagues' structural model holds that the distance between hydrogen atoms is 2.19 ångströms. This is surprisingly large – almost twice that of other hydrides exhibiting high-temperature superconductivity[16–18] – and suggests that there is relatively little hydrogen present in the authors' samples compared with in similar superconducting compounds.

If the nitrogen doping is indeed partly responsible for the superconducting state, its role in achieving such a high transition temperature is yet to be determined. Further research will be needed to confirm that Dasenbrock-Gammon and co-workers' material is a high-temperature superconductor, and then to understand whether this state is driven by vibration-induced Cooper pairs – or by an unconventional mechanism that is yet to be uncovered.

Regardless of the mechanism, the prospect of a material that superconducts under ambient conditions is tantalizing. Superconducting materials make powerful magnets that are used, for example, in magnetic resonance imaging (MRI) – a technology that has had a profound impact on medical diagnostics since it first appeared half a century ago[19]. Such magnets can also be used to levitate objects, and this ability has inspired the idea of a high-speed train that would optimize energy consumption by minimizing friction.

But standard MRI systems currently require expensive refrigeration in the absence of high-temperature superconducting components. And trains that travel as fast as aircraft on a fraction of the power are still a thing of the future. Perhaps Dasenbrock-Gammon and colleagues' hydride compound will bring us a step closer to such technologies becoming a reality.

ChangQing Jin is at the Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China. **David Ceperley** is in the Department of Physics, University of Illinois Urbana-Champaign, Urbana, Illinois 61801, USA.
e-mails: Jin@iphy.ac.cn; ceperley@illinois.edu

1. Chu, C. W. *Physica C* **482**, 33–44 (2012).
2. Drozdov, A. P., Eremets, M. I., Troyan, I. A., Ksenofontov, V. & Shylin, S. I. *Nature* **525**, 73–76 (2015).
3. Dasenbrock-Gammon, N. *et al. Nature* **615**, 244–250 (2023).
4. Bardeen, J., Cooper, L. N. & Schrieffer, J. R. *Phys. Rev.* **108**, 1175–1204 (1957).
5. Ashcroft, N. W. *Phys. Rev. Lett.* **21**, 1748–1749 (1968).
6. McMahon, J. M., Morales, M. A., Pierleoni, C. & Ceperley, D. M. *Rev. Mod. Phys.* **84**, 1607–1653 (2012).
7. Ashcroft, N. W. *Phys. Rev. Lett.* **92**, 187002 (2004).
8. Drozdov, A. P. *et al. Nature* **569**, 528–531 (2019).
9. Somayazulu, M. *et al. Phys. Rev. Lett.* **122**, 027001 (2019).
10. Li, Z. *et al. Nature Commun.* **13**, 2863 (2022).
11. Ma, L. *et al. Phys. Rev. Lett.* **128**, 167001 (2022).
12. Keimer, B., Kivelson, S. A., Norman, M. R., Uchida, S. & Zaanen, J. *Nature* **518**, 179–186 (2015).
13. Ishida, K., Nakai, Y. & Hosono, H. *J. Phys. Soc. Jpn* **78**, 62001 (2009).
14. Snider, E. *et al. Nature* **586**, 373–377 (2020).
15. Snider, E. *et al. Nature* **610**, 804 (2022).
16. Duan, D. F. *et al. Sci. Rep.* **4**, 6968 (2014).
17. Peng, F. *et al. Phys. Rev. Lett.* **119**, 107001 (2017).
18. Wang, H., Tse, J. S., Tanaka, K., Iitaka, T. & Ma, Y. *Proc. Natl Acad. Sci. USA* **109**, 6463–6466 (2012).
19. Lauterbur, P. C. *Nature* **242**, 190–191 (1973).
20. Gao, G. *et al. Mater. Today Phys.* **21**, 100546 (2021).

The authors declare no competing interests.

Plant sciences

A protein entry path into chloroplasts

Takashi Hirashima & Toshiya Endo

Structures of the machinery for importing proteins into chloroplast organelles of algae, determined using cryo-electron microscopy, have opened a new chapter in efforts to understand how chloroplasts are built. **See p.349**

Chloroplasts are organelles, found in plants and algae, that house hundreds of different proteins responsible for photosynthesis, a process essential for life on Earth. Most chloroplast proteins are encoded by the nuclear genome, are made in the cytosol and contain a chloroplast-targeting signal. To be imported into the chloroplast, such proteins must cross the outer and inner envelope membranes (OEM and IEM, respectively) that surround the organelle. Protein complexes on the OEM (called TOC) and IEM (called TIC) form the intricate machinery that mediates protein transport across the two membranes[1–3]. On page 349, Liu *et al.*[4] report the cryo-electron microscopy (cryo-EM) structures of the TOC–TIC supercomplex from the unicellular green alga *Chlamydomonas reinhardtii*. This and another such structure reported in *Cell* by Jin *et al.*[5] are the first high-resolution structures of the TOC and TIC complexes – they provide surprises and insights, but also raise many questions.

Chloroplast-destined proteins are recognized and move from the cytosol across the OEM with the aid of the TOC complex. Previous

© 2023 Springer Nature Limited. All rights reserved.

biochemical characterization of mainly plant chloroplasts indicates that the TOC complex contains two proteins, Toc34 and Toc159, which function as surface receptors for chloroplast-targeting signals, and the protein Toc75, which forms a structure called a β-barrel and is thought to be a protein-import channel.

The first surprise from the cryo-EM structures is that Toc75 assembles together with another protein, Toc90 (also known as Toc120), which is a *Chlamydomonas* version of the plant protein Toc159, to form a large hybrid transmembrane (TM) β-barrel in the OEM that provides a route for protein entry. This barrel has 30 β-strand segments, and one 'seam' on the side of the barrel is a tightly sealed antiparallel β-sheet, whereas the other seam is formed through weak hydrophobic side-chain contacts. This latter seam might function as a flexible 'lateral gate' to enable the transport of proteins that either insert into the OEM or are too large to fit in the central pore and resist being unfolded for entry into the channel[6].

Another peculiar and unexpected observation is that a green-alga-specific OEM protein, Ctap4 (also known as Toc39), forms a second membrane-embedded β-barrel, which is approximately 12 ångströms from the main part of the TOC complex. The role of this smaller barrel remains to be determined.

Chloroplast proteins that have moved through the TOC complex travel through the aqueous intermembrane space (IMS) and cross the IEM through the TIC complex. This occurs in cooperation with the import motor proteins of the AAA–ATPase protein complex and/or soluble chaperone proteins inside the chloroplast[7]. The import motor might exert a pulling force to drive protein entry; however, this motor was not identified as being associated with the TIC structure determined in the current studies.

A remarkable feature of the TOC–TIC supercomplex is that the TOC and TIC complexes form an interlocked architecture in the IMS, tightly integrating the two complexes (Fig. 1). In both chloroplasts and mitochondria (another organelle surrounded by two membranes), efficient protein entry across the outer and inner membranes requires a coordinated operation. In plant chloroplasts, the TOC and TIC complexes form a stable supercomplex only in the presence of translocating proteins and/or with the aid of the bridging protein Tic236 (ref. 8). However, a relatively stable TOC–TIC supercomplex forms in the absence of these proteins in *C. reinhardtii* chloroplasts[9]. The supercomplex is stable without translocating proteins, and whether it has a version of a bridging protein was unknown. The unveiled cryo-EM structures show that the chloroplast-encoded Tic214 protein and other Tic proteins co-fold to build a unique IMS structure and associate

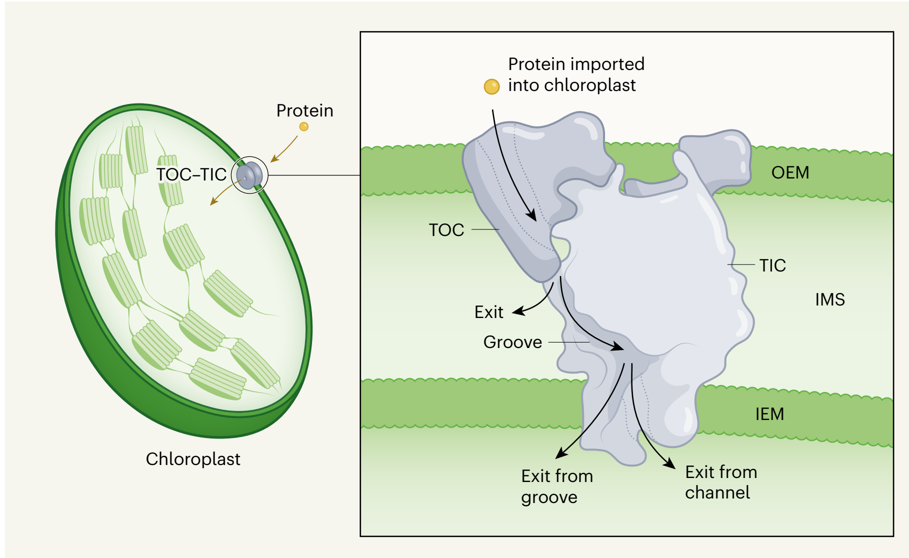


Figure 1 | The TOC–TIC supercomplex from an alga. Liu *et al.*[4] and Jin *et al.*[5] present structures, obtained using cryo-electron microscopy (cryo-EM), of this supercomplex, which is needed for protein entry into chloroplast organelles. The overall structure has components in the outer and inner envelope membrane (OEM and IEM, respectively) and in the intermembrane space (IMS). Some proposed routes for protein movement through a central channel or surface grooves are indicated. The papers differ regarding routes in the IEM-embedded TIC parts (the routes shown correspond to those proposed by Liu and colleagues).

with the IMS domains of the TOC components. In addition, a segment of Tic214 extends even further to enter the TOC hybrid barrel from the IMS side. However, the way in which Tic214 – made inside the chloroplast – is incorporated into the intricate IMS structure is a mystery.

The tight association of the TOC and TIC complexes would create a continuous protein-transport route from the cytosol all the way to inside the chloroplast. Such a route might prevent highly hydrophobic membrane proteins from forming misfolded aggregates in the IMS. However, it might also present a challenge for transporting some soluble proteins to the IMS. To overcome this problem, soluble IMS proteins might exit through the side openings of the TOC hybrid barrel towards the IMS, whereas proteins targeted to the interior of the chloroplast might be directed along the surface groove of the IMS architecture and then reach the protein-transport route in the membrane-embedded part of the TIC complex.

In contrast to the TOC complex, the membrane-embedded part of the TIC complex is organized with approximately 15 TM helices provided by Tic20, Tic214 and other Tic proteins. There is a long-standing debate about whether Tic20 and Tic110 have a central role in protein entry[3], and Tic110 was absent from the TIC structures. This raises the question of how protein entry is organized in the TIC complex. The answer is not clear, and different routes are proposed in the two papers. The TM helices from Tim20 and Tim214 collectively form a narrow conduit in the membrane-embedded portion of the TIC complex. Proteins could potentially move through this channel if it undergoes a conformational change to expand at the site of constriction. Alternatively, proteins might move along the surface groove of the TIC complex that faces the IEM.

In conjunction with mass spectrometry analyses[9], the cryo-EM structures of the TOC–TIC supercomplex have uncovered previously unknown components that are mostly specific to green algae. However, some potential components are missing from the cryo-EM structures, and the structures also contain segments that have yet to be identified. The components of the TOC–TIC systems, including their import motors, exhibit variation in their subunit composition during the development of plastids (organelles that can form structures such as chloroplasts), and have undergone changes during the evolution of algae and land plants, with the exception of their core components[3]. Therefore, gaining a deeper understanding of the precise functions of each component and obtaining high-resolution structures of the TOC–TIC system from land plants, as well as the TIC-associated motor complexes, will add more chapters to our understanding of chloroplast formation.

Takashi Hirashima and **Toshiya Endo** are in the Faculty of Life Sciences and at the Institute for Protein Dynamics, Kyoto Sangyo University, Kyoto 603-8555, Japan.
e-mail: tendo@cc.kyoto-su.ac.jp

1. Schleiff, E. & Becker, T. *Nature Rev. Mol. Cell Biol.* **12**, 48–59 (2011).
2. Jarvis, P. & López-Juez, E. *Nature Rev. Mol. Cell Biol.* **14**, 787–802 (2013).
3. Nakai, M. *Biochim. Biophys. Acta Bioenerg.*

© 2023 Springer Nature Limited. All rights reserved.

1847, 957–967 (2015).
4. Liu, H., Li, A., Rochaix, J.-D. & Liu, Z. *Nature* **615**, 349–357 (2023).
5. Jin, Z. *et al. Cell* **185**, 4788–4800 (2022).
6. Ganesan, I., Shi, L.-X, Labs, M. & Theg, S. M. *Plant Cell* **30**, 2161–2173 (2018).
7. Kikuchi, S. *et al. Plant Cell* **30**, 2677–2703 (2018).
8. Chen, Y.-L. *et al. Nature* **564**, 125–129 (2018).
9. Ramundo, S. *et al. Proc. Natl Acad. Sci. USA* **117**, 32739–32749 (2020).

The authors declare no competing interests.
This article was published online on 28 February 2023.

Cancer

Fatty acids prime the lung as a site for tumour spread

Laura V. Pinheiro & Kathryn E. Wellen

The mechanisms that enable the deadly spread of cancer are not fully understood. It emerges that tumours can signal to the lung to manipulate lipids and so prime the organ to support tumour cells that subsequently spread there.

Cancer cells require nutrients such as glucose, amino acids and fatty acids to grow and proliferate. They also require nutrients to disperse and grow in organs distant from the initial (primary) tumour site – a spreading process called metastasis. The nutritional environment affects whether cancer cells can thrive in a new location. Tumour cells are known to secrete factors that act on distant sites ahead of their arrival, to prepare a favourable environment, termed a pre-metastatic niche[1]. However, nutritional aspects of the pre-metastatic niche are poorly understood. Writing in *Nature Cancer*, Altea-Manzano *et al.*[2] reveal how breast cancer cells program the metabolic environment in the lung before colonization, to support metastatic growth.

Fatty acids are a class of nutrient implicated in supporting metastasis, and high-fat diets promote metastasis in mice[3,4]. However, the factors that affect fatty-acid availability at sites of metastasis are not well known. Altea-Manzano *et al.* find that two fatty acids in particular, palmitate and oleate, are highly abundant in the extracellular fluid of the healthy lung and liver – common sites of metastasis. Feeding mice a high-fat diet further increases the fatty-acid abundance in these tissues, and coincides with increased metastatic growth in a breast cancer model.

The authors then considered a crucial question. Could tumour-derived signals remodel the metabolic environment of distant tissues to improve the ability of cancer cells to grow there? The researchers tested this by injecting mice with factors secreted by tumours *in vitro*. Remarkably, this treatment resulted in a rise in palmitate in the lung extracellular fluid, indicating that signals from a distant tumour can change the nutritional environment in the lung. Lung extracellular fluid collected post-mortem from people with breast cancer had higher levels of palmitate than that of individuals who had not had cancer, supporting the relevance of these findings for humans.

A cell type called a lung-resident alveolar type II (AT2) cell was identified as the source of palmitate-containing lipids (Fig. 1a). AT2 cells produce lung surfactant, a complex of phospholipids and proteins that functions to reduce surface tension at the air–liquid interface in the lung's alveolar structures, preventing their collapse. Surfactant is rich in lipids, suggesting that tumours might exploit the normal metabolic function of the AT2 cells.

To understand how palmitate gives metastasizing cancer cells an advantage, Altea-Manzano *et al.* used an experimental system to model features of tumours *in vivo* and to mimic 3D tumour growth. The set-up consisted of a culture system in which cells aggregate and grow into sphere-like structures[5]. Palmitate, but no other fatty acids tested, increased spheroid growth in a manner that required lipid breakdown to generate the metabolic intermediate acetyl coenzyme A (acetyl-CoA). Acetyl-CoA is needed for a protein modification called acetylation, which can regulate protein function, and the abundance of acetyl-CoA can affect the acetylation of certain proteins, including some that are involved in regulating gene expression[6].

Altea-Manzano and colleagues found that acetyl-CoA derived from palmitate is used by an enzyme called KAT2A to acetylate subunit p65 of the transcription-factor protein NF-κB. This acetylation leads to increased expression of genes that promote metastasis. Although several nutrient sources can supply acetyl-CoA, palmitate, but no other nutrients tested, increased the expression of KAT2A, possibly accounting for the role of palmitate, but not other fatty acids, in this setting.

Targeting various components of this pathway in the cancer cells, including palmitate breakdown and KAT2A, potently suppressed spheroid growth in culture and tumour metastasis in mice. These findings delineate a mechanism whereby signals from the tumour instruct AT2 cells to produce and secrete lipids, increasing palmitate abundance in the pre-metastatic niche (Fig. 1b). Palmitate is then taken up by the metastasizing cancer

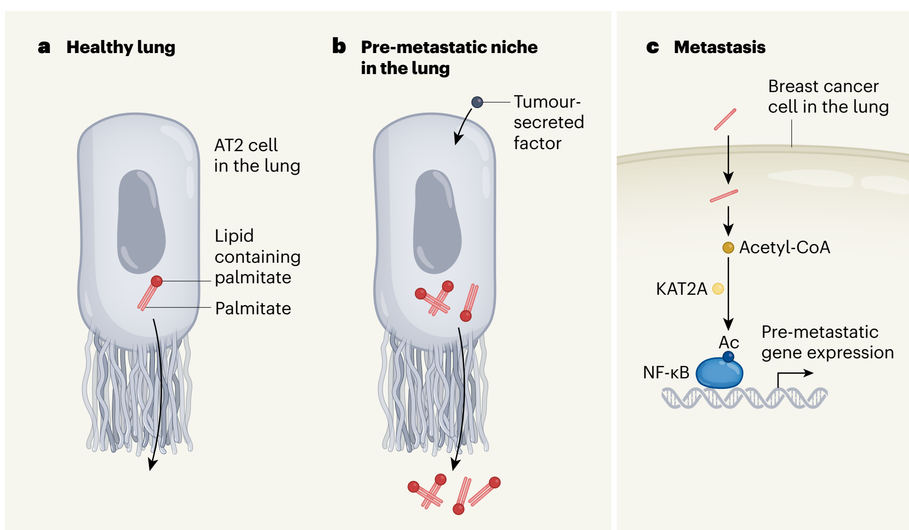


Figure 1 | Tumours manipulate lipids to create a favourable environment for tumour spread. Altea-Manzano *et al.*[2] investigated how breast cancer cells spread to the lungs in mice. **a**, AT2 cells in the lungs synthesize and secrete lipid molecules that contain the fatty acid palmitate. **b**, The authors report that unknown tumour-secreted factors boost palmitate production. This creates a favourable site for tumour spread (termed a pre-metastatic niche). **c**, When breast cancer cells themselves reach the lungs (through a process called metastasis), they take up palmitate and convert it to acetyl-CoA. This molecule is used by the enzyme KAT2A to add an acetyl group (Ac) to the transcription-factor protein NF-κB. NF-κB regulates gene expression to support tumour growth.

© 2023 Springer Nature Limited. All rights reserved.

cells and used to activate a gene-expression program that allows successful growth in the lung (Fig. 1c).

Several intriguing questions arise from this work. In terms of understanding the mechanism involved, a key unknown is the identity of the tumour-cell signal that acts on AT2 cells. Identifying this signal and discovering how it increases the secretion of palmitate-containing lipids in AT2 cells would elucidate key details of how tumour cells nutritionally prime the pre-metastatic niche, and potentially present opportunities for therapeutic intervention.

Lung surfactant consists mainly of phospholipids, a major component of which is the molecule dipalmitoylphosphatidylcholine, which contains two acyl chains derived from palmitate[7]. Indeed, palmitate in the extracellular fluid is mainly a component of larger lipid molecules. Whether the cancer cells selectively take up specific lipids from the lung environment, and if so, how, remains to be determined. In addition, it will be worth investigating the mechanism by which palmitate specifically regulates KAT2A expression.

Another avenue of research prompted by this work is the interplay between dietary fat, lipids in the pre-metastatic niche and metastasis. The authors found that feeding mice a high-fat diet increases the number of AT2 cells; thus, diet might increase palmitate in the pre-metastatic niche through mechanisms beyond simply providing the fatty acid. With that in mind, it would be interesting to test whether this effect is also specific to palmitate, or whether other fatty acids can similarly modulate the number of AT2 cells. Notably, another study found that dietary palmitate, but not the fatty acids oleate or linoleate, boosts the initiation of metastasis through a reprogramming mechanism (an epigenetic modification) that acts on the complex of DNA and protein called chromatin[8]. It is appealing to imagine future strategies that combine dietary interventions with targeted therapies to suppress metastatic growth.

Finally, because the mechanisms described in Altea-Manzano and colleagues' study use the surfactant-producing function of AT2 cells and thus are specific to the lung environment, it will be important to discover whether distinct mechanisms enable metabolic adaptations for metastasis to other sites. Breast cancer metastasis to the brain requires increased palmitate synthesis by cancer cells there, consistent with the idea that different sites require distinct adaptations[9]. Given the common requirement for palmitate for metastasis to both lung and brain, the data raise the question of whether targeting fatty-acid synthesis throughout the body would decrease metastasis to multiple sites. Inhibitors of fatty-acid synthesis are currently being tested in clinical trials for the treatment of primary tumours and some metastatic cancers[10].

Altea-Manzano and colleagues' work reveals key aspects of how tumour cells facilitate colonization of the lung. The pathway identified might offer new molecular targets for suppressing metastasis — a major challenge for the treatment of cancer.

Laura V. Pinheiro and **Kathryn E. Wellen** are in the Department of Cancer Biology and at the Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.
e-mail: wellenk@upenn.edu

1. Peinado, H. *et al. Nature Rev. Cancer* **17**, 302–317 (2017).
2. Altea-Manzano, P. *et al. Nature Cancer* https://doi.org/10.1038/s43018-023-00513-2 (2023).
3. Broadfield, L. A., Pane, A. A., Talebi, A., Swinnen, J. V. & Fendt. S.-M. *Dev. Cell* **56**, 1363–1393 (2021).
4. Martin-Perez, M., Urdiroz-Urricelqui, U., Bigas, C. & Benitah, S. A. *Cell Metab.* **34**, 1675–1699 (2022).
5. Elia, I. *et al. Nature* **568**, 117–121 (2019).
6. Sivanand, S., Viney, I. & Wellen, K. E. *Trends Biochem. Sci.* **43**, 61–74 (2018).
7. Agudelo, C. W., Samaha, G. & Garcia-Arcos, I. *Lipids Health Dis.* **19**, 122 (2020).
8. Pascual, G. *et al. Nature* **599**, 485–490 (2021).
9. Ferraro, G. B. *et al. Nature Cancer* **2**, 414–428 (2021).
10. Batchuluun, B., Pinkosky, S. L. & Steinberg, G. R. *Nature Rev. Drug Discov*. **21**, 283–305 (2022).

The authors declare no competing interests.
This article was published online on 28 February 2023.

scientific reports



Publish your research in psychology

You will now find a brand-new section of *Scientific Reports* content dedicated specifically to psychology, with the aim of addressing the psychology community's specific needs.

- **Expert** Editorial Board to manage your paper
- Follows Nature Portfolio's **high peer review standards**
- Indexed in **Web of Science, PubMed** and other major repositories
- Research accessed from over **180 countries worldwide**

nature.com/srep

A116069

nature portfolio

© 2023 Springer Nature Limited. All rights reserved.

Deuterium-enriched water ties planet-forming disks to comets and protostars

https://doi.org/10.1038/s41586-022-05676-z

Received: 2 June 2022

Accepted: 21 December 2022

Published online: 8 March 2023

Check for updates

John J. Tobin[1]✉, Merel L. R. van 't Hoff[2], Margot Leemker[3], Ewine F. van Dishoeck[3], Teresa Paneque-Carreño[3,4], Kenji Furuya[5], Daniel Harsono[6], Magnus V. Persson[7], L. Ilsedore Cleeves[8], Patrick D. Sheehan[9] & Lucas Cieza[10,11]

Water is a fundamental molecule in the star and planet formation process, essential for catalysing the growth of solid material and the formation of planetesimals within disks[1,2]. However, the water snowline and the HDO:H_2O ratio within proto-planetary disks have not been well characterized because water only sublimates at roughly 160 K (ref. [3]), meaning that most water is frozen out onto dust grains and that the water snowline radii are less than 10 AU (astronomical units)[4,5]. The sun-like protostar V883 Ori ($M_* = 1.3\ M_\odot$)[6] is undergoing an accretion burst[7], increasing its luminosity to roughly 200 $L_\odot$ (ref. [8]), and previous observations suggested that its water snowline is 40–120 AU in radius[6,9,10]. Here we report the direct detection of gas phase water (HDO and $H_2{}^{18}O$) from the disk of V883 Ori. We measure a midplane water snowline radius of approximately 80 AU, comparable to the scale of the Kuiper Belt, and detect water out to a radius of roughly 160 AU. We then measure the HDO:H_2O ratio of the disk to be $(2.26 \pm 0.63) \times 10^{-3}$. This ratio is comparable to those of protostellar envelopes and comets, and exceeds that of Earth's oceans by 3.1σ. We conclude that disks directly inherit water from the star-forming cloud and this water becomes incorporated into large icy bodies, such as comets, without substantial chemical alteration.

英文杂志全球首发QQ群:855583774

We observed V883 Ori (located in the Orion molecular clouds) using the Atacama large millimetre/submillimetre array (ALMA) at roughly 0.1″ resolution (40 AU diameter at its distance of 400 pc; ref. [11]). The full extent of the proto-planetary disk surrounding V883 Ori is well-resolved in our observations, with the dust and gas emission extending to roughly 125 and 320 AU, respectively[6]. The disk around V883 Ori is young, similar to the well-known HL Tau system[12], and the disk mass of roughly 0.02–0.09 $M_\odot$ is about ten times greater than the surrounding envelope mass[6,13]. The accretion burst (which may have begun more than 130 years ago[7,14]) has heated a large fraction of the disk above the sublimation temperature of water, and we illustrate how V883 Ori and a more typical proto-planetary disk differ in Fig. 1.

We detect three emission lines of gas phase water in the disk of V883 Ori: HDO at 225.89672 and 241.561550 GHz and $H_2{}^{18}O$ at 203.40752 GHz (Methods). The HDO and $H_2{}^{18}O$ kinematics are consistent with Keplerian rotation (Extended Data Figs. 1 and 2). The emission from all three lines peak at a radius of 50 to 60 AU, but extend to roughly 160 AU (Fig. 2 and Extended Data Fig. 3). HDO is substantially brighter than $H_2{}^{18}O$ (despite the D:H ratio less than 1) owing to the ratio of ^{16}O:^{18}O and its molecular properties (Methods). The structure and extent of the water emission is very similar to complex organic molecules (COMs), such as methanol[9,15], which makes sense given that methanol is expected to sublimate at a similar temperature (roughly 130 K)[3]. $C^{17}O$, however, is more extended than water and methanol because it sublimates at around 25 K (ref. [16]) (Fig. 2). The apparent depression of emission in the centre of some images in Fig. 2 is due to the opacity of the continuum, absorbing gas emission at radii less than 40 AU. The high dust optical depth at smaller radii, even for more-evolved disks[17], may obscure molecular line emission coming from the disk midplane and make the midplane water emission of lower luminosity systems difficult to characterize.

The warm nature of V883 Ori's disk enables us to characterize its water reservoir with spatially resolved observations, unlike most proto-planetary disks (Fig. 1)[18]. We measured the column densities of HDO and $H_2{}^{18}O$ in several ways to test the robustness of our measurements (Methods). For our primary method, we extracted the integrated spectra within a 0.4″ (160 AU) radius to measure the total flux of the emission lines (Extended Data Fig. 4). We also extracted radial intensity profiles from the integrated intensity maps in Fig. 2 using Keplerian masks. The spectral extraction method yields a disk-averaged measurement, whereas the integrated intensity method yields the measurements as a function of disk radius. The integrated fluxes of the HDO lines are 0.644 ± 0.028 and 0.595 ± 0.037 Jy km s^{-1} for the 225 and 241 GHz lines, respectively, whereas the $H_2{}^{18}O$ line flux is 0.126 ± 0.025 Jy km s^{-1}.

[1]National Radio Astronomy Observatory, Charlottesville, VA, USA. [2]Department of Astronomy, University of Michigan, Ann Arbor, MI, USA. [3]Leiden Observatory, Leiden University, Leiden, The Netherlands. [4]European Southern Observatory, Garching, Germany. [5]National Astronomical Observatory of Japan, Mitaka, Japan. [6]Institute of Astronomy, Department of Physics, National Tsing Hua University, Hsinchu, Taiwan. [7]Department of Space, Earth and Environment, Chalmers University of Technology, Onsala Space Observatory, Onsala, Sweden. [8]Department of Astronomy, University of Virginia, Charlottesville, VA, USA. [9]Center for Interdisciplinary Exploration and Research in Astronomy, Northwestern University, Evanston, IL, USA. [10]Núcleo de Astronomía, Facultad de Ingeniería y Ciencias, Universidad Diego Portales, Santiago, Chile. [11]Millennium Nucleus on Young Exoplanets and their Moons (YEMS), Universidad Diego Portales, Santiago, Chile. ✉e-mail: jtobin@nrao.edu

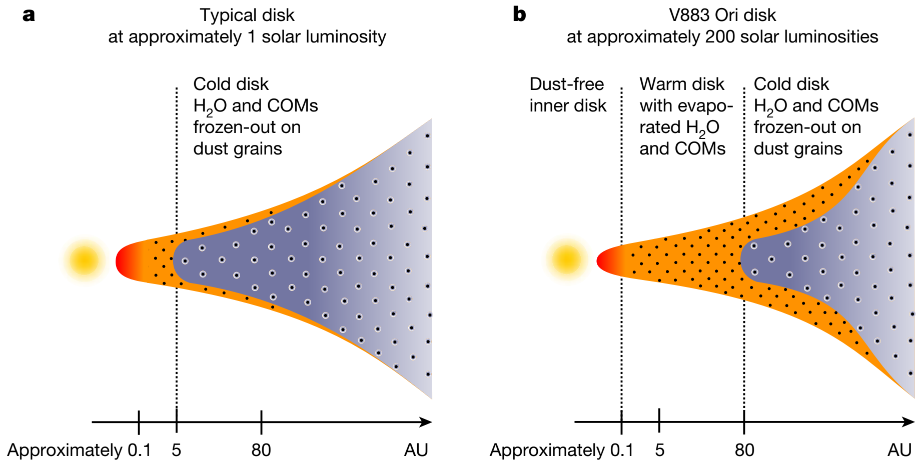


Fig. 1 | Illustrations comparing a typical proto-planetary disk and the disk surrounding the outbursting protostar V883 Ori. a, A typical disk in which water and COMs are frozen out onto dust grains at most radii except the very inner disk. **b**, The case of V883 Ori in which the disk has a much larger region where water and COMs are sublimated from the dust grains, enabling the water and COM emission to be detected and resolved. Orange and red regions denote the warmer regions of the disk where the ices are sublimated and the light blue/grey denotes the colder regions where the water and COMs are frozen out. The black points alone denote dust grains with no ice coating, whereas the black points with grey circles around them denote icy dust grains.

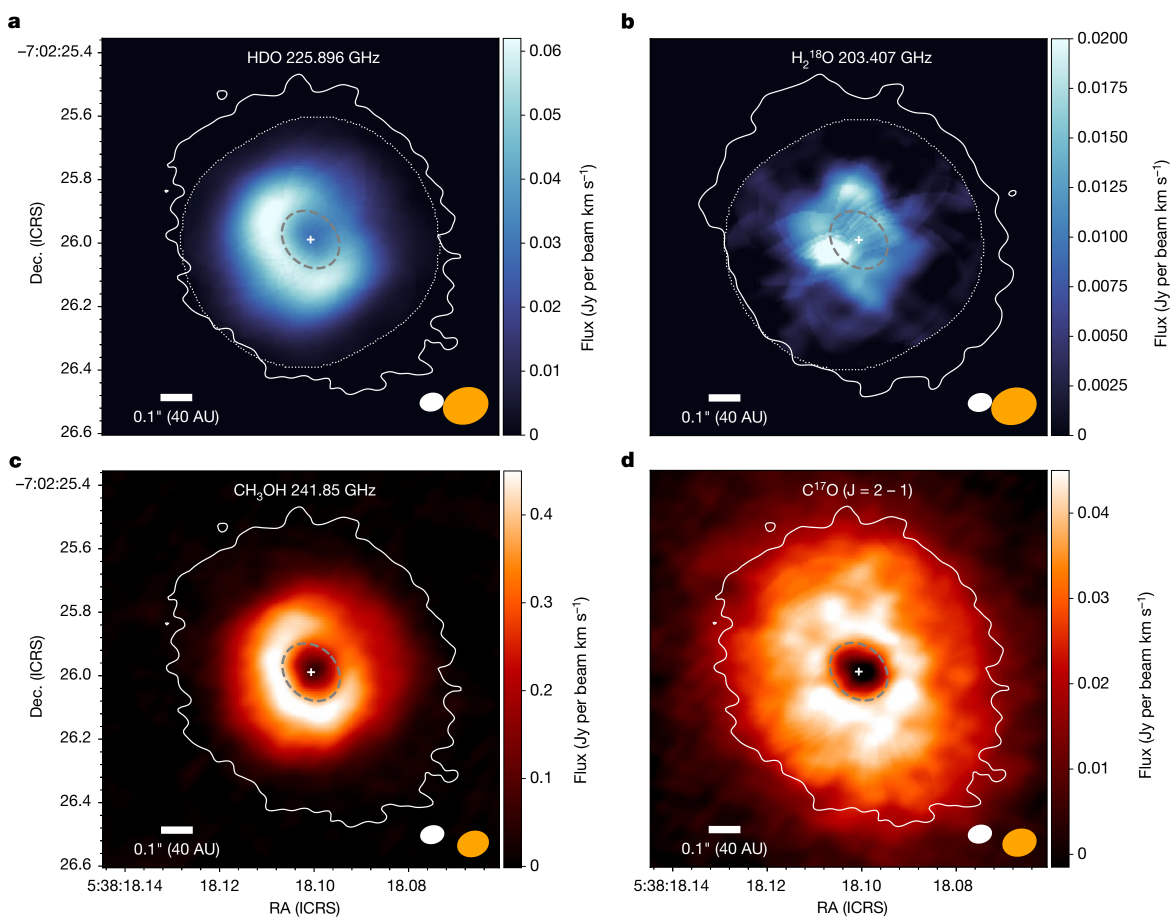


Fig. 2 | Integrated intensity images of water and other molecular lines in the disk of V883 Ori. a–d, We show the HDO 225 GHz line (**a**), the $H_2{}^{18}O$ 203 GHz line (**b**), CH_3OH (**c**) and $C^{17}O$ (**d**), as the colour scale, whereas the outer extent of the dust continuum emission is shown as a white contour; the integrated intensity maps were extracted from the data using the Keplerian masks whose outer extent are marked as dotted lines in **a**, **b**. The white cross marks the location of the continuum emission peak and the position of the protostar. The HDO and $H_2{}^{18}O$ lines show emission that is smaller in radial extent than the continuum and $C^{17}O$ emission. The CH_3OH (methanol) image shows a very similar structure and extent relative to the HDO and $H_2{}^{18}O$ lines. $C^{17}O$ more fully traces the extent of the disk gas emission with its lower sublimation temperature of roughly 25 K, extending beyond the continuum emission. The central depression in emission for all lines is the result of optically thick continuum emission attenuating the molecular emission at radii smaller than roughly 40 AU (0.1″); the extent of this optically thick region is denoted with the dashed thick grey line in the centre of each image. The depression is less obvious for $H_2{}^{18}O$ in **b** due to its lower signal to noise ratio and some contamination of its integrated intensity map from a neighbouring line. The ellipses in the lower right corner denote the resolution of the line observations (orange, roughly 0.1″) and the continuum (white, roughly 0.08″). Dec., declination; ICRS, International Celestial Reference System; RA, right ascension.

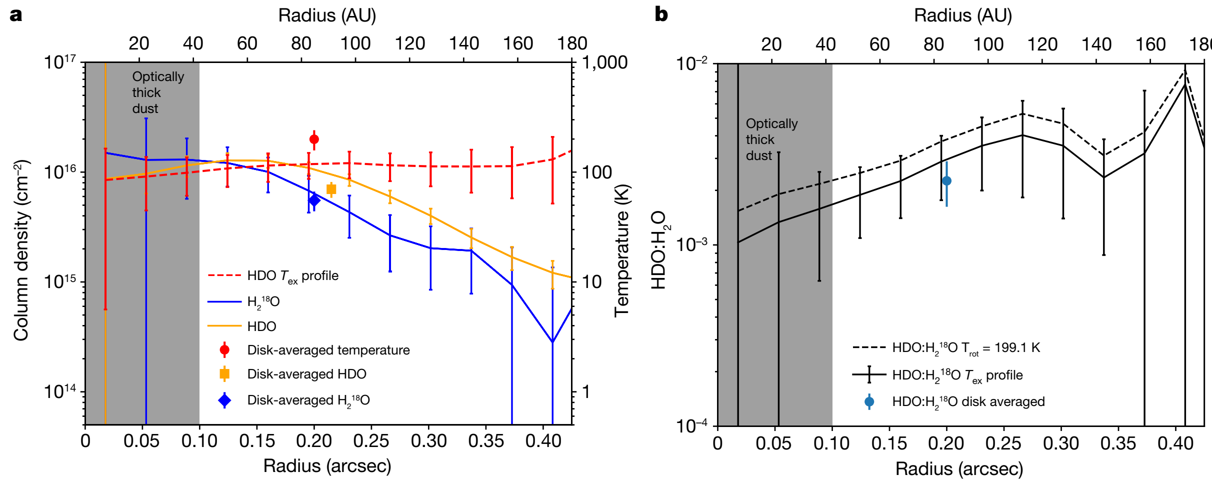


Fig. 3 | Radial column density measurements of water and the HDO:H_2O ratio along with their disk-averaged measurements. a,b, The radial column density profiles of HDO and $H_2^{18}O$ (**a**) and the HDO:H_2O ratio as a function of deprojected radius in the disk of V883 Ori (**b**). The disk-averaged values are plotted at 0.2″ in both panels. The HDO column density is the average of the column density derived from the two HDO transitions (this only reduces the uncertainties because the HDO column density will be identical for the two transitions). The HDO column densities have some contamination from a neighbouring COM line for the 241 GHz transition, which has been quantified in the Methods. This contamination could affect the temperature and the HDO:H_2O profile, but the overall effect is small as shown by the agreement with the disk-averaged value, which has has the contamination of the HDO 241 GHz line removed. The same excitation temperature profile derived from the HDO radial intensity profiles (**a**) (red dashed line) is used to calculate both the HDO and $H_2^{18}O$ column density profiles (Methods). The HDO:H_2O ratio calculated with a fixed, 199 K, excitation temperature (T_{ex}) is shown as the dashed line and the solid line uses the HDO excitation temperature profile. The error bars are only shown on the solid line for clarity, but are similar for a constant excitation temperature. The uncertainties drawn correspond to 1 s.d. in the measurements from standard error propagation. We shade the radius less than 0.1″ (40 AU) where the disk is optically thick and the measurements are unreliable.

The ratio of the upper level column densities for the two HDO lines is used to estimate the water excitation temperature (199 ± 42 K). This temperature is used to derive the column densities for HDO and $H_2^{18}O$ of $(6.98 \pm 1.15) \times 10^{15}$ and $(5.52 \pm 1.13) \times 10^{15}$ cm^{-2}, respectively. The $H_2^{18}O$ column density translates to a H_2O column density of $(3.09 \pm 0.70) \times 10^{18}$ cm^{-2} (Methods). The radial column density profiles of HDO and $H_2^{18}O$ are shown in Fig. 3 and are consistent with the disk-averaged measurements. The midplane water snowline is estimated to be roughly 80 AU, corresponding to the radius at which the HDO and $H_2^{18}O$ column density begins to rapidly decline (Fig. 3). The total gaseous water mass in the disk is 1.69×10^{27} g, equivalent to roughly 1,200 Earth oceans. This is a lower limit because it does not include water at radii less than 40 AU or the water ice in the outer disk. Finally, we calculate a disk-averaged HDO:H_2O ratio of $(2.26 \pm 0.63) \times 10^{-3}$, consistent with the HDO:H_2O radial profile (Fig. 3). The measurements of the HDO:H_2O ratios from different methods are all consistent within their 1 s.d. uncertainties, demonstrating that the measurement method does not strongly influence the results (Extended Data Table 4).

Water (H_2O, HDO, and D_2O) is expected to form as ice mantles on dust grains in the cold interstellar medium (ISM) through grain surface reactions. The abundance of deuterated species becomes enhanced in the cold ISM due to deuterium becoming locked into H_2D^+, and the dissociative recombination of deuterium bearing molecules increases the free deuterium available for HDO and D_2O formation within ice mantles[18,19]. The HDO:H_2O ratio in ice mantles is not well-constrained, only upper limits of roughly 10^{-2} are available[20]. However, within the warm gas at small radii around Class 0 protostars, where these ices have sublimated, the HDO:H_2O ratios are found to be between roughly 6×10^{-4} and 2×10^{-3} (refs. [21–23]). Then comets range from 3×10^{-4} to 10^{-3} (refs. [24,25]), and Earth's oceans are 3.11×10^{-4} (Extended Data Table 5). The HDO:H_2O ratios for V883 Ori and these different objects are summarized in Fig. 4.

The HDO:H_2O ratio of V883 Ori is comparable to the youngest (Class 0) protostars[22], and then compared to more-evolved objects that formed within the Solar System, V883 Ori's HDO:H_2O ratio is also similar to many Oort Cloud comets and 67P from the Jupiter Family comets (Fig. 4). This makes sense because comets are icy relics that are expected to have formed in the outer parts of the proto-planetary disk. It is important to note that it was previously thought that the Jupiter Family and Oort cloud comets formed at different locations in the Solar

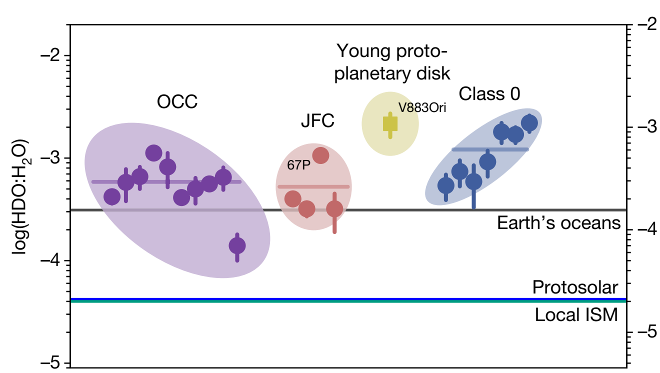


Fig. 4 | HDO:H_2O ratio for Class 0 protostars, V883 Ori, Jupiter Family comets (JFC, with 67P labelled), Oort Cloud comets (OCC), Earth's oceans, the Sun and the local ISM. The points are arranged to highlight a potential evolutionary scenario with the forming protostars on the right, the 'evolved' Solar System bodies on the left, and V883 Ori fills in a region of parameter space for a young proto-planetary disk, just before or contemporaneous with planet formation. The ellipses drawn around the different groups are meant to guide the eye, but the horizontal lines within each ellipse denote the mean HDO:H_2O ratio for each group, and the vertical lines associated with each point represent the 1 s.d. uncertainties. If a vertical line is not visible, the uncertainty is less than the size of the symbol. In the context of other measurements, V883 Ori indicates that the HDO:H_2O ratio does not strongly evolve from the protostar phase to the disk, providing further observational evidence that the water is directly inherited from the envelope to the disk without significant chemical changes. Moreover, the HDO:H_2O ratio of comets are similar to V883 Ori (but slightly below), indicating that the water we probe in the gas phase within the disk of V883 Ori is similar to the water that becomes incorporated into comets. The individual measurements are tabulated in the Methods (Extended Data Table 5).

System, close to the orbits of the giant planets and outside the orbit of Neptune, respectively. It is now thought that the populations are mixed[26], but their HDO:H_2O ratios should reflect their original formation location. However, Earth's oceans and some comets have a lower HDO:H_2O compared to V883 Ori, indicating that the water in those bodies may have undergone further processing at high temperatures, lowering their HDO:H_2O ratios. Or, the HDO:H_2O ratio is intrinsically lower in the bodies responsible for water delivery to Earth.

Indeed, it has been shown that outbursts can lower the HDO:H_2O ratio in the inner 1 to 3 AU of a disk through a combination of high gas temperature (greater than 500 K), radial mixing and isotope exchange reactions in the gas phase ($HDO + H_2 \rightarrow H_2O + HD$)[27]. However, this fast evolution of the gaseous HDO:H_2O ratio is only taking place at radii smaller than we can probe in V883 Ori (and inside the optically thick dust). Conversely, at radii greater than 10 AU, the lower temperature (200–300 K) and densities mean that the isotope exchange reactions are extremely slow; it would take greater than 100 Myr to significantly change the observed HDO:H_2O ratio in V883 Ori[28]. This means that V883 Ori's HDO:H_2O ratio is tracing that of water sublimated from the ices. The slowness of the isotope exchange reactions and the fact that HDO is only efficiently created in ice, means that the water emission in V883 Ori, even if it is coming from upper layers of the disk (Fig. 1), is expected to have a similar HDO:H_2O ratio to the disk midplane.

The HDO:H_2O ratio within the disk of V883 Ori, at the end of the protostar phase, fills in a crucial gap in the water trail just before or contemporaneous with the formation of large solid bodies within the disk. The relative constancy of the HDO:H_2O ratio from the protostar phase to the proto-planetary disk indicates that the water within the disk of V883 Ori is directly inherited from the infalling envelope[29–31]. Furthermore, the comet 67P, protostars and V883 Ori[15] all have CH_2DOH:CH_3OH ratios that are roughly ten times their HDO:H_2O ratios. This is further evidence that both water and methanol (and the deuterated species) are formed on icy dust grains, that the molecules are inherited from the prestellar phase, and that significant chemical reset does not occur during disk or comet formation. Some chemical changes are observed from the envelope to disk[32], but the effect is not strong enough to affect the water and probably other COMs that arrive to the disk as ice[33,34]. Although the specific delivery mechanism of water on Earth remains debated (comets and/or asteroids)[35], the high D:H found in V883 Ori is evidence that the water molecules in our Solar System originated in the cold ISM before the formation of the Sun[29]. Therefore, spatially resolved water observations towards young planet-forming disks are crucial in linking the water reservoir and the formation of terrestrial planets.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-05676-z.

1. Blum, J. & Wurm, G. The growth mechanisms of macroscopic bodies in protoplanetary disks. *Annu. Rev. Astron. Astrophys.* **46**, 21–56 (2008).
2. Okuzumi, S., Tanaka, H., Kobayashi, H. & Wada, K. Rapid coagulation of porous dust aggregates outside the snow line: a pathway to successful icy planetesimal formation. *Astrophys. J.* **752**, 106 (2012).
3. Penteado, E. M., Walsh, C. & Cuppen, H. M. Sensitivity analysis of grain surface chemistry to binding energies of ice species. *Astrophys. J.* **844**, 71 (2017).
4. Lecar, M., Podolak, M., Sasselov, D. & Chiang, E. On the location of the snow line in a protoplanetary disk. *Astrophys. J.* **640**, 1115–1118 (2006).
5. Carr, J. S., Najita, J. R. & Salyk, C. Measuring the water snow line in a protoplanetary disk. *Res. Notes. Am. Astron. Soc.* **2**, 169 (2018).
6. Cieza, L. A. et al. Imaging the water snow-line during a protostellar outburst. *Nature* **535**, 258–261 (2016).
7. Strom, K. M. & Strom, S. E. The discovery of two FU orionis objects in L1641. *Astrophys. J. L.* **412**, L63 (1993).
8. Furlan, E. et al. The Herschel Orion protostar survey: spectral energy distributions and fits using a grid of protostellar models. *Astrophys. J. Supp.* **224**, 5 (2016).
9. van 't Hoff, M. L. R. et al. Methanol and its relation to the water snowline in the disk around the young outbursting star V883 Ori. *Astrophys. J. L.* **864**, L23 (2018).
10. Leemker, M. et al. Chemically tracing the water snowline in protoplanetary disks with HCO^+. *Astron. Astrophys.* **646**, A3 (2021).
11. Kounkel, M. et al. The APOGEE-2 survey of the Orion star-forming complex. II. Six-dimensional structure. *Astron. J.* **156**, 84 (2018).
12. ALMA Partnership. The 2014 ALMA long baseline campaign: first results from high angular resolution observations toward the HL Tau region. *Astrophys. J. L.* **808**, L3 (2015).
13. Sheehan, P. D., Tobin, J. J., Looney, L. L. & Megeath S. T. The VLA/ALMA nascent disk and multiplicity (VANDAM) survey of Orion protostars VI. Insights from radiative transfer modeling. *Astrophys. J.* **929**, 76S (2022).
14. Pickering, E. C. Detection of new nebulae by photography. *Ann. Harvard College Observ.* **18**, 113–117 (1890).
15. Lee, J.-E. et al. The ice composition in the disk around V883 Ori revealed by its stellar outburst. *Nat. Astron.* **3**, 314–319 (2019).
16. Collings, M. P. et al. A laboratory survey of the thermal desorption of astrophysically relevant molecules. *Mon. Not. R. Astron. Soc.* **354**, 1133–1140 (2004).
17. Andrews, S. M., Wilner, D. J., Hughes, A. M., Qi, C. & Dullemond, C. P. Protoplanetary disk structures in Ophiuchus. II. Extension to fainter sources. *Astrophys. J.* **723**, 1241–1254 (2010).
18. van Dishoeck, E. F. et al. Water in star-forming regions: physics and chemistry from clouds to disks as probed by Herschel spectroscopy. *Astron. Astrophys.* **648**, A24 (2021).
19. Tielens, A. G. G. M. Surface chemistry of deuterated molecules. *Astron. Astrophys.* **119**, 177–184 (1983).
20. Parise, B. et al. Search for solid HDO in low-mass protostars. *Astron. Astrophys.* **410**, 897–904 (2003).
21. Persson, M. V., Jørgensen, J. K., van Dishoeck, E. F. & Harsono, D. The deuterium fractionation of water on solar-system scales in deeply-embedded low-mass protostars. *Astron. Astrophys.* **563**, A74 (2014).
22. Jensen, S. S. et al. ALMA observations of water deuteration: a physical diagnostic of the formation of protostars. *Astron. Astrophys.* **631**, A25 (2019).
23. Jensen, S. S. et al. ALMA observations of doubly deuterated water: inheritance of water from the prestellar environment. *Astron. Astrophys.* **650**, A172 (2021).
24. Altwegg, K. et al. 67P/Churyumov-Gerasimenko, a Jupiter family comet with a high D/H ratio. *Science* **347**, 1261952 (2015).
25. Altwegg, K., Balsiger, H. & Fuselier, S. A. Cometary chemistry and the origin of icy solar system bodies: the view after Rosetta. *Annu. Rev. Astron. Astrophys.* **57**, 113–155 (2019).
26. Altwegg, K. et al. D_2O and HDS in the coma of 67P/Churyumov-Gerasimenko. *Philos. Trans. R. Soc. London Ser. A.* **375**, 20160253 (2017).
27. Owen, J. E. & Jacquet, E. Astro- and cosmochemical consequences of accretion bursts - I. The D/H ratio of water. *Mon. Not. R. Astron. Soc.* **446**, 3285–3296 (2015).
28. Lécluse, C. & Robert, François Hydrogen isotope exchange reaction rates: origin of water in the inner solar system. *Geochim. Cosmochim. Acta* **58**, 2927–2939 (1994).
29. Cleeves, L. I. et al. The ancient heritage of water ice in the solar system. *Science* **345**, 1590–1593 (2014).
30. Furuya, K., van Dishoeck, E. F. & Aikawa, Y. Reconstructing the history of water ice formation from HDO/H_2O and D_2O/HDO ratios in protostellar cores. *Astron. Astrophys.* **586**, A127 (2016).
31. Furuya, K. et al. Water delivery from cores to disks: deuteration as a probe of the prestellar inheritance of H_2O. *Astron. Astrophys.* **599**, A40 (2017).
32. Sakai, N. et al. Change in the chemical composition of infalling gas forming a disk around a protostar. *Nature* **507**, 78–80 (2014).
33. Visser, R., van Dishoeck, E. F., Doty, S. D. & Dullemond, C. P. The chemical history of molecules in circumstellar disks. I. Ices. *Astron. Astrophys.* **495**, 881–897 (2009).
34. Drozdovskaya, M. N., Walsh, C., Visser, R., Harsono, D. & van Dishoeck, E. F. Methanol along the path from envelope to protoplanetary disc. *Mon. Not. R. Astron. Soc.* **445**, 913–929 (2014).
35. O'Brien, D. P., Izidoro, A., Jacobson, S. A., Raymond, S. N. & Rubie, D. C. The delivery of water during terrestrial planet formation. *Space Sci. Rev.* **214**, 47 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2023

Methods

V883 Ori was observed with ALMA located on llano de Chajnantor in Northern Chile in November 2021. Observations were conducted in band 6 (roughly 1.3 mm) and band 5 (roughly 1.5 mm) to target the HDO and $H_2^{18}O$ spectral lines, respectively.

Band 5 observations

The band 5 observations were conducted with two executions on 3 and 4 November 2021 during C-7 configuration with a maximum baseline length of roughly 3,000 kλ, with precipitable water vapour values of 0.72 and 0.78 mm, respectively. There were 41 antennas operating during the first execution and 43 during the second. Each execution was roughly 80 min in duration with roughly 38 min on source per execution for a total time on source in band 5 of roughly 76 min. The monitored quasar J0538-4405 was the absolute flux density and bandpass calibrator in the first execution and J0423-0120 was the flux density and bandpass calibrator in the second execution. Both executions used J0541-0541 as the phase calibrator. The central frequency of the observations was roughly 195.8 GHz and we simultaneously observed many molecular transitions (Extended Data Table 1), but our main focus in this paper is the $H_2^{18}O$ transition at 203.40752 GHz. The absolute flux density uncertainty in band 5 is expected to be roughly 5%.

Band 6 observations

The band 6 observations were conducted in two consecutive executions on 1 November 2021 during C-7 configuration with a maximum baseline length of roughly 4,000 kλ. The measured precipitable water vapour values were 0.74 and 0.77 mm. There were 48 antennas operating during both executions and each execution was roughly 75 min in duration with roughly 40.5 min on source, yielding a total time on source in band 6 of roughly 81 min. Both executions used J0423-0120 as the absolute flux density and bandpass calibrator and J0541-0541 as the phase calibrator. The absolute flux density uncertainty in band 6 is expected to be roughly 5%. The central frequency of the observations was roughly 233.5 GHz and we simultaneously observed many molecular transitions (Extended Data Table 1), but our main focus in this paper is the two HDO transitions at 225.89672 and 241.5615499 GHz. However, we compare to the methanol (CH_3OH) transitions observed in a separate spectral window with high spectral resolution.

Data reduction

The data were reduced using the ALMA calibration pipeline within CASA v.6.2.1 (ref. [36]). We used the pipeline-calibrated visibility data and the continuum regions identified by the ALMA pipeline task hif_findcont to perform self-calibration on the continuum data and applying the phase and amplitude solutions to the line data as well. Continuum subtraction was also performed using the line-free continuum regions for fitting the underlying continuum using the CASA task uvcontsub.

The HDO line at 225 GHz was imaged with the CASA task tclean using a robust parameter of 2.0, a channel spacing of 0.162 km s^{-1}, an image size of 768 × 768 and a pixel size of 0.0075″. All the images used the modified Briggs weighting scheme, set with the tclean parameter weighting='briggsbwtaper'. The resulting beam was 0.104″ × 0.082″. Then the HDO line at 241 GHz was created using the same image parameters, but with a channel spacing of 0.151 km s^{-1} and a beam of 0.098″ × 0.077″. The $H_2^{18}O$ line at 203 GHz was imaged with the same parameters as the HDO lines, but with a channel spacing of 0.4 km s^{-1} due to the lower signal to noise ratio of the line emission compared to the HDO lines; the beamsize for $H_2^{18}O$ is 0.14″ × 0.11″. The other data cubes, including methanol, were created using the same parameters and their native channel spacing (Extended Data Table 1). To facilitate consistency in the analysis, we generated cubes for each HDO transition that use the same restoring beam as the $H_2^{18}O$ cube and 0.4 km s^{-1} channels. This provides a more consistent comparison of the derived radial profiles for HDO and $H_2^{18}O$ and we use these data with the same channel width and restoring beams exclusively for analysis of the HDO:H_2O ratio. Finally, aggregate continuum images were created for the band 5 and band 6 data individually, using the line-free regions from all the spectral windows in each band combined. These continuum images had the same image and pixel sizes as the cubes. The synthesized beams of the band 5 and 6 continuum images are 0.11″ × 0.088″ and 0.073″ × 0.055″, respectively.

Analysis

Keplerian masks. The presence of other strong emission lines from COMs preclude a simple extraction of spectral line fluxes from the disk of V883 Ori. It is known that $H_2^{18}O$ can be contaminated by nearby CH_3OCH_3 (dimethyl ether) lines, and HDO at 225 GHz can be contaminated by CH_3OCHO (methyl formate)[21,22]. Furthermore, HDO at 241 GHz also has contamination, which is likely to come from CH_3CHO (acetaldehyde). Thus, we use Keplerian masks to isolate the HDO and $H_2^{18}O$ emission as best as possible. Keplerian masks select areas within particular channels of a data cube where emission from a Keplerian disk is expected on the basis of the known mass of the protostar. We make use of a Keplerian mask implementation[37] in which various parameters (for example, stellar mass, disk radius, system velocity, emitting height, line width an so on) can be tuned to optimize the mask fit. However, even the best tuned masks include a small amount of contaminating flux at the expense of better capturing the HDO and $H_2^{18}O$ flux. HDO 241 GHz has the largest amount of contamination from its neighbouring COM line, whereas $H_2^{18}O$ also has some contamination. We show the channel maps of the HDO and $H_2^{18}O$ emission with the Keplerian masks overlaid in Extended Data Figs. 1 and 2. The Keplerian masks for the contaminating lines are also drawn to show the expected velocities and location of the contaminating emission that overlap with the HDO and $H_2^{18}O$ masks.

The Keplerian mask parameters used for each line are provided in Extended Data Table 2; we used consistent parameters between the Keplerian masks for the water lines and other COMs, different parameters were only necessary when the radial extent of emission was different or there were several hyperfine components to the transition, as in the case of $C^{17}O$. The exact same Keplerian mask parameters are used for the two HDO lines and the $H_2^{18}O$ line for consistency in the analysis.

We then make use of these masks to extract moment maps of the emission and construct radial intensity profiles of the HDO and $H_2^{18}O$ emission, using the bettermoments package[38,39]. In addition to the Keplerian masks, we also apply velocity range limits on the moment map calculations to omit unavoidable contaminating flux from neighbouring lines. As such, the 225 GHz HDO integrated intensity map includes velocities between 0.05 and 8.2 km s^{-1} and the 241 GHz HDO and $H_2^{18}O$ maps include velocities between 1.25 and 8.2 km s^{-1}. Comparing with the HDO 225 GHz line, we find that the more limited ranges for HDO at 241 GHz and $H_2^{18}O$ would only omit roughly 2% of the total line flux. The integrated intensity map for HDO at 241 GHz is shown in Extended Data Fig. 3.

The measured line fluxes within the Keplerian masks for HDO at 225 and 241 GHz are 0.554 ± 0.007 and 0.68 ± 0.007 Jy km s^{-1}, respectively. Then the $H_2^{18}O$ line flux was 0.12 ± 0.009 Jy km s^{-1}. The values for HDO 241 GHz and $H_2^{18}O$ are larger than they should be because flux from contaminating spectral lines is present in overlapping velocity channels and falls within the Keplerian masks. If we only integrate the line flux in the velocity range from 5.05 to 7.05 km s^{-1}, where contamination is minimal, we can scale to account for the missing flux by using the HDO 225 GHz line (ratio of the flux between 5.05 and 7.05 km s^{-1} to the total line flux, 0.37 ± 0.008). We measure a scaled HDO 241 GHz line flux of 0.535 ± 0.016 Jy km s^{-1} and a scaled $H_2^{18}O$ line flux of 0.093 ± 0.014 Jy km s^{-1}. The scaled HDO 225 GHz line flux is identical to the

measurement from the full Keplerian mask value by definition. The line fluxes from the Keplerian masks in the full and 5.05 to 7.05 km s^{-1} velocity ranges (scaled and unscaled) are provided in Extended Data Table 4.

We note that there is a small amount of contamination to the HDO 241 GHz line flux within the 5.05 to 7.05 km s^{-1} velocity range. We correct for the estimated contamination to the HDO 241 GHz line flux in this velocity range by subtracting the estimated contamination, which is 2.3% of the total CH_3CHO line flux (0.32 Jy km s^{-1} estimated from spectral extraction, section Spectral extraction). The 5.05 to 7.05 km s^{-1} spectral range for HDO 241 GHz corresponds to 7.33 to 9.33 km s^{-1} from the line centre for CH_3CHO. We then use the ratio of the total HDO 225 GHz line flux between 7.33 and 9.33 km s^{-1} to the total HDO 225 GHz line flux to estimate that HDO at 241 GHz is contaminated by roughly 2.3% of the total CH_3CHO line flux. This method of removing the estimated line flux only works because between 5.05 and 7.05 km s^{-1} the CH_3CHO Keplerian mask is wholly within the HDO 241 GHz Keplerian mask.

The uncertainties on the line fluxes are statistical only and do not include the estimated 5% absolute flux calibration uncertainty.

Moment 0 extraction. While the nearby COM spectral lines to the two HDO and $H_2^{18}O$ lines preclude a standard moment 0 map from accurately measuring the line fluxes, we can make use of a standard moment 0 map within the limited velocity range of 5.05 to 7.05 km s^{-1}. This enables us to measure both the emission morphology and the significance of the line emission for the HDO and $H_2^{18}O$ lines without any potential bias caused by the pixel selections of the Keplerian masks. We show the moment 0 image computed between 5.05 and 7.05 km s^{-1} in Extended Data Fig. 3 and overlay the contours of the $H_2^{18}O$ emission on the HDO emission to demonstrate the consistency of their emission morphologies.

We also measure the line fluxes directly from these moment 0 maps within a circle having a 0.4″ radius, centred on the continuum source. We measure line fluxes of 0.244 ± 0.007, 0.233 ± 0.008 and 0.052 ± 0.008 for HDO 225 GHz, HDO 241 GHz and $H_2^{18}O$, respectively. The HDO 241 GHz line flux is also corrected for the same amount of contamination from CH_3CHO line flux between 5.05 and 7.05 km s^{-1} as we did when the flux was measured within a Keplerian mask in the section on Keplerian masks. The line fluxes are also reported in Extended Data Table 4. We do notice that the line fluxes from the standard moment 0 extraction are systematically larger than those from the Keplerian masks in the same velocity range. The ratio of the Keplerian mask measured line fluxes between 5.05 and 7.05 km s^{-1} to those from the moment 0 measurements are 0.84 ± 0.03, 0.88 ± 0.03 and 0.65 ± 0.14 for HDO 225 GHz, HDO 241 GHz and $H_2^{18}O$, respectively. This small discrepancy could result from the Keplerian mask excluding some line flux, but as we show later this does not lead to a significant difference in the measurement of the column densities of HDO and $H_2^{18}O$ nor the HDO:H_2O ratio. Moreover, the scaled line fluxes from the moment 0 analysis are consistent within the 1σ uncertainties of the line fluxes from spectral extraction (Keplerian masks).

Spectral extraction. In addition to the analysis of the data cubes, we also perform an analysis of the disk-averaged spectra. We extract the spectra around the HDO, $H_2^{18}O$, and methanol transitions within a 0.4″ radius circle centred on the continuum emission peak. The input data cubes are first converted from Jy beam^{-1} to Jy pix^{-1} and then all emission contained within the 0.4″ radius of the protostar position is summed, per channel, to create a one-dimensional spectrum. The spectra around the HDO and $H_2^{18}O$ lines are shown in Extended Data Fig. 4.

We also made use of spectral stacking, which removes the Keplerian kinematic pattern from the data and aligns the emitting components to be at the same velocity using the spectral-cube package[40,41]. This method boosts the signal to noise ratio of the spectral lines and makes the features sharper by lowering their velocity widths because of the removal of the double-peaked line shape that is associated with Keplerian rotation. The spectrally stacked data make contaminating line features easier to identify and associate with the rest frequencies of molecular species. However, we do not perform measurements with the spectrally stacked data because a forward model using the spectral template is expected to be more reliable. Moreover, the spectral stacking can affect the line profiles in unexpected ways if not all emission lines come from the same emission layers within the disk or the midplane. Therefore, it is advantageous to use a spectral template to forward model the raw spectra. The stacked spectral data are presented in Extended Data Fig. 4 and clearly show that $H_2^{18}O$ is indeed detected with a peak at its expected frequency.

Spectral fitting. We constructed two different template spectra to disentangle the HDO and $H_2^{18}O$ from the contaminating lines. First, because the integrated intensity maps for methanol, HDO and $H_2^{18}O$ show that these molecules all have very similar spatial and velocity distributions, we used the isolated methanol lines that we detect in another spectral window as a spectral template for HDO, $H_2^{18}O$ and other COM lines in the disk. Second, we constructed a toy Keplerian disk model with a central protostar mass of 1.3 $M_\odot$ (ref. [6]) using the Line Modelling Engine (LIME)[42] radiative transfer code to simulate the expected line profile of an optically thin molecule that is present in the disk between a radius of 20 and 85 AU. The model also excludes dust emission to make as ideal a spectrum as possible.

To construct the template spectrum using the methanol lines, we first extracted the five isolated methanol lines in our spectral window that targeted methanol at high spectral resolution (Extended Data Table 1), using the same aperture radius (0.4″) used to extract the other spectra. We then baseline subtracted the individual spectra to take out any slope in the spectral baseline. We next normalized each spectrum to have a maximum value of 1.0. Following normalization, we upsampled the spectral resolution of these lines by a factor of six and put all the individual methanol spectra on the same frequency axis. The spectra were averaged together to make a higher signal to noise template spectrum. We again baseline subtracted the new template spectrum to take out a residual slope in the spectral baselines. Finally, we mirrored the spectrum and averaged the mirror with itself to create a symmetric spectral profile. The mirroring creates a more idealized line profile for modelling emission lines other than methanol that may not have the exact same line shape.

To extract the one-dimensional spectrum from the LIME model, we use the same routines that we used to extract the spectra from the real data cubes and the 0.4″ extraction radius encompasses all the emission in the model. We also mirror this spectrum to smooth out any asymmetries that remain from the model gridding and noise in the radiative transfer calculations. The LIME and methanol template spectra are shown in Extended Data Fig. 5. The wings of the template lines agree well, and the primary difference is in the central portion of the line and to a lesser extent the location of the symmetric peaks. This difference is due to the high opacity of the methanol lines, which causes the centre of their line profiles to appear more filled-in relative to the peaks at higher and lower frequencies. These template spectra enable us to deblend the HDO and $H_2^{18}O$ lines from the COM lines without using the spectrally stacked data. Both spectral templates are resampled to the lower spectral resolution of the data for fitting.

The HDO lines were all brighter than the nearby contaminating COM lines, and each HDO line was blended with roughly two COM lines. Generally, a single line was the primary contaminating feature and there was tentative contamination from a second COM line. Then $H_2^{18}O$ had contamination primarily from one nearby COM line, but two other lines of the same species are also nearby. We list the most likely contaminating COM lines in Extended Data Table 3. The plausibility of the contaminating features relied on the positive identification of the species towards V883 Ori from previous work[15] and that the feature

could be present with a reasonable column density (10^{14}–10^{16} cm^{-2}) and a temperature of 200–300 K, as simulated with the eXtended CASA Line Analysis Software Suite (XCLASS)[43]. Similar to previous work, we find that $H_2^{18}O$ has contamination from CH_3OCH_3, HDO 225 GHz has contamination from CH_3OCHO (refs. [21,22]), but we also find that HDO at 225 GHz could be contaminated by $^{13}CH_3CHO$ (an acetalaldehyde isotopologue). Then, HDO at 241 GHz was found to be most probably contaminated by CH_3CHO, with a possible second, weaker contaminating feature from $(CH_3)_2CO$ (acetone).

With the frequencies of the contaminating lines identified, and the well-constrained velocity of the source (4.25 km s^{-1}) determined from the high spectral resolution methanol emission detected in our data set, we are able to model the spectral features of HDO and $H_2^{18}O$ and their surrounding lines with a linear addition of the scaled template spectra. We provided the rest frequencies for each line and used the scipy function minimize to find the best fitting multiplicative factors to fit the spectra using scaled spectral templates; rest frequencies of the spectral features were kept fixed. The results of the spectral fitting are shown in Extended Data Fig. 4.

We decided to adopt the optically thin spectral template for our fiducial model because it qualitatively fits the HDO spectra better than the methanol model. However, the subtle differences in the line fluxes whether we use the optically thin model, the methanol model, or the optically thin model for HDO and $H_2^{18}O$ and the methanol model for COMs does affect the excitation temperature, and hence the line column densities and HDO:H_2O ratios. However, the HDO:H_2O ratios are all consistent within their 1σ uncertainties, so the choice of model does not significantly affect the results as a whole.

The line fluxes were measured from the spectrum with the contaminating COM lines subtracted and uncertainties are derived from the root mean square of the residual spectrum with all fitted lines removed. This approach takes into account any uncertainties associated with imperfect fitting of the contaminating lines. We also tested Markov chain Monte Carlo methods to fit the spectra, but the results did not differ from the maximum likelihood fit using scipy'sminimize function. However, to validate our assumption of a fixed velocity and not allow the source velocity to be optimized, we randomly sampled velocities within 5% of 4.25 km s^{-1} and determined whether the width of the distribution of fitted line flux densities was smaller or larger than the statistical uncertainty from the root mean square of the residuals. For all the water lines, the statistical uncertainty was smaller than the error in flux that could result from an imperfect system velocity. The line fluxes derived from the spectral template models are given in Extended Data Table 4.

The measurements of the line fluxes for all spectral template models are not entirely in agreement between the spectral fitting methods, Keplerian mask extraction methods and scaled moment 0 methods. Some differences can be explained by contamination of the HDO 241 GHz line in particular, but the largest differences are for the HDO 225 GHz line flux from the Keplerian mask and the scaled moment 0 line flux from the 5.05 to 7.05 km s^{-1} moment 0 map. But, most flux differences are within 3σ, and the with HDO 241 GHz and $H_2^{18}O$ being consistent within 2σ. Some difference can be expected for several reasons. First, we integrate the spectra over the same circular aperture in all channels, whereas the Keplerian mask has more limited spatial extent because it only covers the portion of the disk expected to be emitting at a particular velocity. However, the projected outer extent of the Keplerian masks is comparable to 0.4″ spectral extraction radius. Second, the masks are also more limited in their spectral extents to avoid too much contamination from other lines, particularly for HDO at 241 GHz and $H_2^{18}O$. Third, some true line emission can extend beyond the Keplerian mask at a given velocity channel and this would be specifically excluded from the Keplerian mask line flux, but included in the integrated spectrum. The most easily quantified difference is the difference in spectral range of the Keplerian mask, which results in just a 2% flux difference. The impact of these differences on the resultant column densities are explored further in the section 'HDO:H_2O results from Keplerian mask and moment 0 line fluxes'.

Measuring column densities. We measure the column densities of HDO and $H_2^{18}O$ following standard methods under the assumption that the lines are optically thin and in local thermodynamic equilibrium (LTE)[44]. The spectral line parameters and partition functions were taken from the NASA Jet Propulsion Laboratory (JPL) database[45]. We first measure the line integrated intensities in units of Jy km s^{-1}, which we then convert to temperature units through

$$\int T_B \, dv = \frac{c^2}{2\nu^2 k_B \Omega} \int F_\nu \, dv, \quad (1)$$

where k_B is the Boltzmann constant, T_B is the brightness temperature, F_ν the flux density, ν the frequency, c the speed of light, $\Omega = \pi R^2$, and R the 0.4″ aperture radius (converted to radians) used for extracting the spectra. The radial profiles are extracted from the data cubes in K km s^{-1} using the bettermoments package[38,39].

The column densities of the upper energy level are determined from the equation

$$N_u = \frac{8\pi k_B \nu_u^2}{hc^3 A_{ul}} \int T_B \, dv \quad (2)$$

where h is Planck's constant, ν is the frequency of the transition, and A_{ul} is the Einstein A coefficient for the transition. An estimate of the excitation temperature is then needed to determine total column densities, which we obtain from the ratio of column densities in the HDO 225 and 241 GHz lines through

$$T_{ex} = \frac{E_{u,225\,GHz} - E_{u,241\,GHz}}{\ln\left(\frac{N_{u,225\,GHz}}{N_{u,241\,GHz}} \frac{g_{u,241\,GHz}}{g_{u,225\,GHz}}\right)} \quad (3)$$

where E_u is the upper level energy of the two HDO transitions in units of Kelvin and g_u is the statistical weight for each transition. Then the total column density for each transition is given by

$$N = \frac{N_u}{g_u} Q_{rot}(T_{ex}) e^{\frac{E_u}{T_{ex}}} \quad (4)$$

where $Q_{rot}(T_{ex})$ is the partition function and T_{ex} is the excitation temperature. We use the tabulated partition functions from the JPL database as a function of temperature for HDO and $H_2^{18}O$, and for T_{ex} values in between the tabulated temperatures we interpolate. Also, we use the full partition function for $H_2^{18}O$ that includes both the *ortho* and *para* states, implicitly correcting for the *ortho* to *para* ratio (OPR) of $H_2^{18}O$. The calculated values for the interpolated partition functions are listed in Extended Data Table 4. The OPR for $H_2^{18}O$ is assumed to be three, substantial differences from this amount are not expected. The comet 67P was found to have an OPR of 2.94 ± 0.06 (ref. [46]), which is consistent with the laboratory work finding that evaporating ices have an OPR of three[47]. We convert the $H_2^{18}O$ column density to an H_2O column density by multiplying by the ^{16}O:^{18}O ratio of 560 ± 25 (ref. [48]). However, we do know that the ^{16}O:^{18}O in 67P for H_2O is 445 ± 35 (ref. [49]). Therefore, it is possible that the actual H_2O column densities in V883 Ori are lower, which would further elevate the HDO:H_2O ratio.

The uncertainties in the column densities and ratios are derived from standard error propagation. The uncertainty on the HDO column densities includes the random measurement error from the noise in each channel, the roughly 5% absolute flux density calibration uncertainty (only included once in the average HDO column density as both lines are observed simultaneously), and the uncertainty in the excitation temperature. The $H_2^{18}O$ also includes the further uncertainty on the

ratio of ^{16}O to ^{18}O (neglecting the possibility of a large systematic uncertainty). These uncertainties are all propagated into the HDO:H_2O ratio such that the uncertainties on our derived ratio include all the known uncertainties and better reflect the absolute accuracy of the measurement versus only including statistical measurement uncertainties. Finally, we did not make a correction for extra dust attenuation of the 225 and 241 GHz HDO lines relative to $H_2{}^{18}O$ at 203 GHz when calculating the HDO:H_2O ratio. This correction would have a radial dependence and would depend on the adopted power-law dust opacity dependence with frequency. The application of such a correction would only result in a higher HDO column density and a higher HDO:H_2O ratio as dust attenuation is larger at higher frequencies.

We further note that the excitation temperature derived is not expected to be very accurate because of the small range of upper level energies sampled by the HDO lines (95 and 167 K), but the HDO:H_2O ratio only varies by a factor of roughly 2 for temperatures between 75 and 225 K, so the precision of the excitation temperature is not extremely important.

The assumption of LTE is appropriate for the V883 Ori disk. The critical densities of the HDO 225, 241, and $H_2{}^{18}O$ 203 GHz are 9.72×10^5, 1.06×10^6 and 3.79×10^5 cm^{-3}, respectively, using the collision coefficients for a temperature of roughly 200 K from the Leiden Atomic and Molecular Database (coefficients for *para*-H_2O are used for *para*-$H_2{}^{18}O$)[50,51]. These critical densities are satisfied even for protostars in which the origin of the emission is likely to have contributions from the envelope outside the disk. Moreover, previous analyses have examined the difference between LTE and radiative transfer models that take into account non-LTE effects[21,22]. Those studies find that the HDO:H_2O ratios derived from non-LTE modelling are consistent with the LTE calculations or differ by factors of only 3 to 4. Also, they often indicate even higher HDO:H_2O ratios and are thus not regarded as being more accurate. Finally, a simple estimate of the disk density in the upper layers at a height of 50 AU and a radius of 80 AU we find that the density would be roughly 7.4×10^6 cm^{-3}. All other emission would be from lower layers at higher density and also fulfilling the conditions for LTE. We estimated the disk density by assuming parameters typical for an irradiated disk using a parameterized surface and vertical density profile[52]

$$\rho(r,z) = \frac{\Sigma(r)}{\sqrt{2\pi}H(r)}\exp\left(-\frac{z^2}{2H(r)^2}\right). \quad (5)$$

where r is the cylindrical radius, z is the height above the midplane, and $\Sigma(r)$ is

$$\Sigma(r) = \Sigma_0\left(\frac{r}{10\ \mathrm{AU}}\right)^{-1} \quad (6)$$

and $\Sigma_0 = 100$ g cm^{-2}, as is typical for a roughly 0.01 $M_\odot$ proto-planetary disk[17], chosen as a conservative lower limit. $H(r)$ for the gas disk is parameterized as

$$H = 1.0\ \mathrm{AU}\left(\frac{r}{10\ \mathrm{AU}}\right)^{1.25} \quad (7)$$

which is typical for an irradiated disk.

HDO:H_2O results from Keplerian mask and moment 0 line fluxes. We noted earlier that the line fluxes extracted using a Keplerian mask and integrated intensity maps from 5.05 to 7.05 km s^{-1} differed from those derived from the spectrum and it is worthwhile examining how line flux extraction using the Keplerian mask would affect our results. Note that we correct for the 2% reductions to the HDO 241 GHz and $H_2{}^{18}O$ line fluxes due to the smaller channel ranges used relative to HDO 225 GHz in their line flux measurements in the calculations that follow. The values for the line fluxes used to calculate the column densities mentioned in the text below are also provided in Extended Data Table 4, in addition to the HDO:H_2O ratios.

We examined line flux measurements that used the full Keplerian mask, which includes some unavoidable contamination from COMs in the 241 GHz and $H_2{}^{18}O$ channel maps. The contamination is difficult to avoid when using an identical Keplerian mask to the HDO 225 GHz line. Because of the contaminated line fluxes to the HDO 241 GHz line, the lower excitation HDO transition (Extended Data Table 3), the excitation temperature is found to be lowest for the full Keplerian mask measurement at 112 ± 3 K. Then the HDO and $H_2{}^{18}O$ column densities are measured to be $(5.07 \pm 0.26) \times 10^{15}$ and $(4.75 \pm 0.44) \times 10^{15}$ cm^{-2}, respectively. These measurements are both lower than those from the spectral extraction, but are consistent at the 1.4 and 0.5σ levels, respectively. The inferred H_2O column density is $(2.66 \pm 0.28) \times 10^{18}$ cm^{-2}, which is corrected for the ^{16}O:^{18}O ratio and the OPR of three. The HDO and H_2O column densities then yield a HDO:H_2O ratio of $(1.91 \pm 0.22) \times 10^{-3}$, which is consistent with the value from spectral extraction within the uncertainties.

Because of the known significant contamination to HDO 241 GHz and $H_2{}^{18}O$ using the same Keplerian mask as HDO 225 GHz, without further optimization, we also examined the measurements obtained by integrating the flux density within a mostly contamination free region of the data cube (5.05 to 7.05 km s^{-1}) and scaling the line flux from that spectral range to the expected flux from the full spectral range using the ratio of HDO 225 GHz flux integrated within a Keplerian mask to the HDO 225 GHz flux present between 5.05 and 7.05 km s^{-1} (0.37 ± 0.008), also within a Keplerian mask. The HDO:H_2O ratio remains unchanged whether we only compare line fluxes between 5.05 and 7.05 km s^{-1} or scale to the expected full line flux, but to compare the column densities consistently, we need to make use of the scaled line fluxes.

The HDO, $H_2{}^{18}O$, and inferred H_2O column densities derived from scaled line fluxes measured within a Keplerian mask from 5.05 to 7.05 km s^{-1} are $(5.36 \pm 0.35) \times 10^{15}$, $(3.76 \pm 0.58) \times 10^{15}$ and $(2.11 \pm 0.34) \times 10^{18}$ cm^{-2}, respectively, using the calculated excitation temperature of 162 ± 11 K. These measurements are both lower than those from the spectral extraction, but are consistent at the 1.1 and 1σ levels, respectively. The HDO and H_2O column densities then yield a HDO:H_2O ratio of $(2.54 \pm 0.44) \times 10^{-3}$, which is consistent within the uncertainties with the value from spectral extraction.

Then, if we instead derive the column densities using the line fluxes measured from the standard moment 0 image computed between 5.05 and 7.05 km s^{-1}, we measure HDO, $H_2{}^{18}O$ and inferred H_2O column densities of $(6.78 \pm 0.62) \times 10^{15}$, $(5.92 \pm 1.00) \times 10^{15}$, $(3.31 \pm 0.59) \times 10^{18}$ cm^{-2} and, respectively, using the calculated excitation temperature of 182 ± 19 K. These values are consistent with the values from spectral extraction within the uncertainties. The HDO and H_2O column densities then yield a HDO:H_2O ratio of $(2.05 \pm 0.41) \times 10^{-3}$, which is consistent within the uncertainties with the value from spectral extraction.

The high degree of consistency of the methods of spectral extraction, standard Keplerian mask, Keplerian mask between 5.05 and 7.05 km s^{-1} and a standard moment 0 between 5.05 and 7.05 km s^{-1} indicates that several methods arrive at compatible measurements for the HDO and $H_2{}^{18}O$ column densities and consistent measurements of the HDO:H_2O ratio. Although the line contamination present in the case of the standard Keplerian mask of the full velocity range is not ideal, it also does not significantly alter the main result. Thus, we conclude that the method of line flux measurement does not significantly affect our results on the column densities and HDO:H_2O ratio measured. The Keplerian mask and standard moment 0 measurements do, however, tend to have lower uncertainties because they make use of measurements over smaller ranges of velocity and/or a smaller number of pixels (through the mask), which reduces the noise that is added from the summation of several channels. We still favour the spectral analysis method because this makes use of the full spectral range of the data rather than subsets of the full data set.

Birth environment of V883 Ori relative to the Sun and other protostars

V883 Ori is compared with Class 0 protostars, comets and water in Earth's oceans in Fig. 4 attempting to determine whether there is an evolutionary sequence in the HDO:H_2O ratios. However, we already know that V883 Ori has formed in a different environment compared to the Sun/Solar System and other Class 0 protostars. V883 Ori is forming within the Orion A giant molecular cloud and is roughly 10 pc from the main cluster/nebula and the nearest protostar is roughly 0.4 pc away: both measurements are in projected distances, thus it is relatively isolated within Orion. By contrast, the Class 0 protostars with measured HDO:H_2O ratios come from a mix of isolated and protostars within small clusters[22] and the isolated Class 0 protostars have systematically higher HDO:H_2O ratios. V883 Ori's HDO:H_2O ratio is most similar to the isolated Class 0 protostars. This may make sense because it is relatively isolated itself, but the isolated protostars are forming in truly isolated cores and are not part of a larger molecular cloud, unlike V883 Ori. It is still unknown what causes the higher HDO:H_2O ratios for isolated systems, but it remains to be seen whether the difference between isolated and clustered protostars remains with larger samples.

The Sun and Solar System were probably formed in a cluster within a giant molecular cloud, which could have been similar to Orion[53,54]. However, the dynamical constraints that are available for the evolution of the Solar System point to its formation environment being closer to, or within the main cluster[55]. Thus, V883 Ori is not forming in completely analogous conditions relative to the early Solar System. Nonetheless, it is still extremely relevant to compare the measurements for V883 Ori to comets within the Solar System and Earth given that they represent the only available measurements of more-evolved bodies. Furthermore, if water is inherited relatively unchanged from the cold molecular cloud phase, the fact that both the Sun/Solar System and V883 Ori formed within giant molecular clouds suggests that it is reasonable to compare the HDO:H_2O ratios of these systems.

Measurements of deuterium to hydrogen ratios from the literature

To create Fig. 4, we collected various published measurements for protostars and comets from the literature. We have assembled these measurements in Extended Data Table 5 for reference.

Data availability

The images used for analysis in the paper are available in the Harvard Dataverse repository (https://doi.org/10.7910/DVN/MDQJEU), along with the reduction scripts used to process the ALMA visibility data and create images. Owing to their size, the raw (and ALMA-pipeline-calibrated) visibility data are only available from the ALMA science archive (https://almascience.nrao.edu/aq/).

Code availability

Codes used for analysis are available in the Harvard Dataverse repository (https://doi.org/10.7910/DVN/MDQJEU), along with the ALMA images used for analysis. Documentation and requirements for various parts of the code are documented.

36. McMullin, J. P., Waters, B., Schiebel, D., Young, W. & Golap, K. CASA architecture and applications. In *Astronomical Data Analysis Software and Systems XVI* Vol. 376, *Proc. Astronomical Society of the Pacific Conference Series* (eds Shaw, R. A. et al.) 127 (Astron. Soc. Pacific, 2007).
37. Teague, R. richteague/keplerian_mask: initial release. *Zenodo* https://doi.org/10.5281/zenodo.4321137 (2020).
38. Teague, R. & Foreman-Mackey, D. A robust method to measure centroids of spectral lines. *Rese. Notes Am. Astron. Soc.* **2**, 173 (2018).
39. Teague, R. Statistical uncertainties in moment maps of line emission. *Res. Notes Am. Astron. Soc.* **3**, 74 (2019).
40. Yen, H.-W. et al. Stacking spectra in protoplanetary disks: detecting intensity profiles from hidden molecular lines in HD 163296. *Astrophys. J.* **832**, 204 (2016).
41. Ginsburg, A., Bally, J., Goddi, C., Plambeck, R. & Wright, M. A Keplerian disk around Orion SrcI, a ~15 $M_\odot$ YSO. *Astrophys. J.* **860**, 119 (2018).
42. Brinch, C. & Hogerheijde, M. R. LIME - a flexible, non-LTE line excitation and radiation transfer method for millimeter and far-infrared wavelengths. *Astron. Astrophys.* **523**, A25 (2010).
43. Möller, T., Endres, C. & Schilke, P. eXtended CASA line analysis software suite (XCLASS). *Astron. Astrophys.* **598**, A7 (2017).
44. Goldsmith, P. F. & Langer, W. D. Population diagram analysis of molecular line emission. *Astrophys. J.* **517**, 209–225 (1999).
45. Pickett, H. M. et al. Submillimeter, millimeter and microwave spectral line catalog. *J. Quant. Spectrosc. Radiat. Transf.* **60**, 883–890 (1998).
46. Cheng, Y. C. et al. Water ortho-to-para ratio in the coma of comet 67P/Churyumov-Gerasimenko. *Astron. Astrophys.* **663**, A43 (2022).
47. Hama, T., Kouchi, A. & Watanabe, N. Statistical *ortho*-to-*para* ratio of water desorbed from ice at 10 Kelvin. *Science* **351**, 65–67 (2016).
48. Wilson, T. L. & Rood, R. Abundances in the interstellar medium. *Annu. Rev. Astron. Astrophys.* **32**, 191–226 (1994).
49. Altwegg, K. et al. Molecule-dependent oxygen isotopic ratios in the coma of comet 67P/Churyumov-Gerasimenko. *Mon. Not. R. Astron. Soc.* **498**, 5855–5862 (2020).
50. Schöier, F. L., van der Tak, F. F. S., van Dishoeck, E. F. & Black, J. H. An atomic and molecular database for analysis of submillimetre line observations. *Astron. Astrophys.* **432**, 369–379 (2005).
51. Faure, A., Wiesenfeld, L., Scribano, Y. & Ceccarelli, C. Rotational excitation of mono- and doubly-deuterated water by hydrogen molecules. *Mon. Not. R. Astron. Soc.* **420**, 699–704 (2012).
52. Williams, J. P. & Cieza, L. A. Protoplanetary disks and their evolution. *Annu. Rev. Astron. Astrophys.* **49**, 67–117 (2011).
53. Adams, F. C. The birth environment of the Solar System. *Annu. Rev. Astron. Astrophys.* **48**, 47–85 (2010).
54. Desch, S. J., Young, E. D., Dunham, E. T., Fujimoto, Y. & Dunlap, D. R. Short-lived radionuclides in meteorites and the Sun's birth environment. Preprint at https://arxiv.org/abs/2203.11169 (2022).
55. Pfalzner, S. & Vincke, K. Cradle(s) of the Sun. *Astrophys. J.* **897**, 60 (2020).
56. Robitaille, T. & Bressert, E. APLpy: astronomical plotting library in Python. Astrophysics Source Code Library, record ascl:1208.017 (ASCL, 2012).
57. Astropy Collaboration et al. The Astropy Project: building an open-science project and status of the v2.0 core package. *Astron. J.* **156**, 123 (2018).
58. Greenfield, P. et al. Astropy: community Python library for astronomy (ASCL, 2013).
59. Thyng, K. M., Greene, C. A., Hetland, R. D., Zimmerle, H. M. & DiMarco, S. F. True colors of oceanography: guidelines for effective and accurate colormap selection. *Oceanography* **29**, 9–13 (2016).
60. Hagemann, R., Nief, G. & Roth, E. Absolute isotopic scale for deuterium analysis of natural waters. absolute D/H ratio for smow. *Tellus* **22**, 712–715 (1970).
61. de Laeter, J. R. et al. Atomic weights of the elements. Review 2000 (IUPAC technical report). *Pure Appl. Chem.* **75**, 683–800 (2003).
62. Brown, R. H., Lauretta, D. S., Schmidt, B. & Moores, J. Experimental and theoretical simulations of ice sublimation with implications for the chemical, isotopic, and physical evolution of icy objects. *Planetary Space Sci.* **60**, 166–180 (2012).
63. Bockelée-Morvan, D. et al. Deuterated water in comet C/1996 B2 (Hyakutake) and its implications for the origin of comets. *Icarus* **133**, 147–162 (1998).
64. Meier, R. et al. A determination of the HDO/H_2O ratio in comet C/1995 O1 (Hale-Bopp). *Science* **279**, 842 (1998).
65. Gibb, E. L. et al. Chemical composition of comet C/2007 N3 (Lulin): another 'atypical' comet. *Astrophys. J.* **750**, 102 (2012).
66. Villanueva, G. L. et al. A sensitive search for deuterated water in comet 8p/Tuttle. *Astrophys. J. L.* **690**, L5–L9 (2009).
67. Bockelée-Morvan, D. et al. Herschel measurements of the D/H and $^{16}O/^{18}O$ ratios in water in the Oort-cloud comet C/2009 P1 (Garradd). *Astron. Astrophys.* **544**, L15 (2012).
68. Hutsemékers, D., Manfroid, J., Jehin, E., Zucconi, J. M. & Arpigny, C. The $^{16}OH/^{18}OH$ and OD/OH isotope ratios in comet C/2002 T7 (LINEAR). *Astron. Astrophys.* **490**, L31–L34 (2008).
69. Biver, N. et al. Radio wavelength molecular observations of comets C/1999 T1 (McNaught-Hartley), C/2001 A2 (LINEAR), C/2000 WM_1 (LINEAR) and 153P/Ikeya-Zhang. *Astron. Astrophys.* **449**, 1255–1270 (2006).
70. Biver, N. et al. Isotopic ratios of H, C, N, O, and S in comets C/2012 F6 (Lemmon) and C/2014 Q2 (Lovejoy). *Astron. Astrophys.* **589**, A78 (2016).
71. Lis, D. C. et al. A Herschel study of D/H in water in the Jupiter-family comet 45P/Honda-Mrkos-Pajdušáková and prospects for D/H measurements with CCAT. *Astrophys. J. L.* **774**, L3 (2013).
72. Hartogh, P. et al. Ocean-like water in the Jupiter-family comet 103P/Hartley 2. *Nature* **478**, 218–220 (2011).
73. Lis, D. C. et al. Terrestrial deuterium-to-hydrogen ratio in water in hyperactive comets. *Astron. Astrophys.* **625**, L5 (2019).

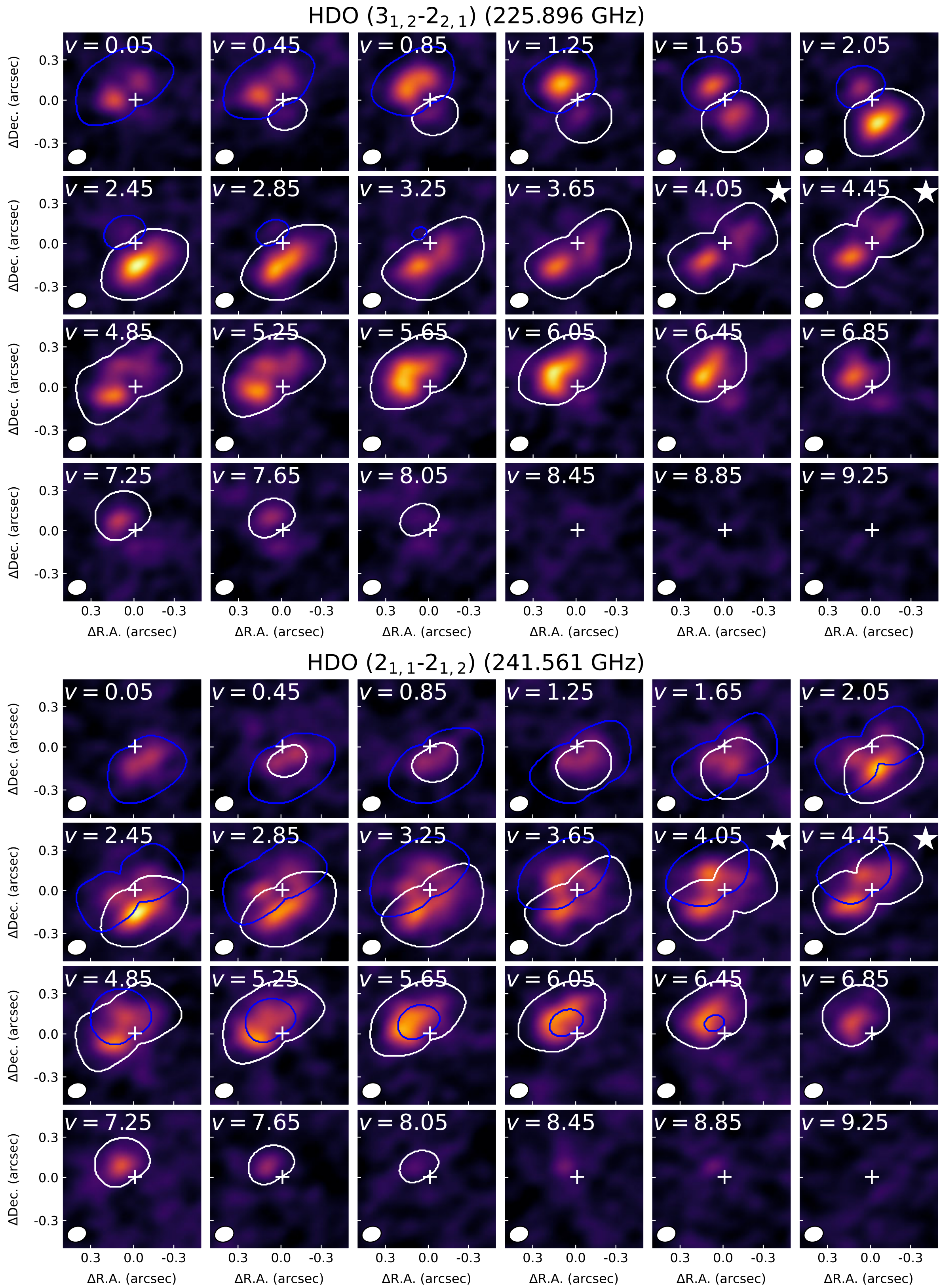
Acknowledgements This paper makes use of the following ALMA data: ADS/JAO. ALMA#2021.1.00186.S. ALMA is a partnership of ESO (representing its member states), National Science Federation (NSF) (USA) and NINS (Japan), together with NRC (Canada), MOST and ASIAA (Taiwan), and KASI (Republic of Korea), in cooperation with the Republic of Chile. The Joint ALMA Observatory is operated by ESO, AUI/NRAO and NAOJ. The National Radio Astronomy Observatory is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc. This research made use of APLpy, an open-source plotting package for Python[56], Astropy (http://www.astropy.org), a community-developed core Python package for Astronomy[57,58], the Python package spectral-cube (https://github.com/radio-astro-tools/spectral-cube) and matplotlib colour

maps from the Python package cmocean[59]. J.J.T. acknowledges support from the National Radio Astronomy Observatory and NASA XRP 80NSSC22K1159. M.L.R.v.H. acknowledges support from the University of Michigan Society of Fellows. M.L. acknowledges support from the Dutch Research Council (NWO) grant no. 618.000.001. E.F.v.D. acknowledges support the NWO, EU A-ERC grant no. 101019751 MOLDISK and the Danish National Research Foundation 'InterCat' grant (no. DNRF150). T.P.-C. acknowledges support from the European Southern Observatory. P.D.S. acknowledges support from NSF grant no. AST-2001830. D.H. is supported by Centre for Informatics and Computation in Astronomy and grant no. 110J0353I9 from the Ministry of Education of Taiwan. D.H. acknowledges support from the Ministry of Science of Technology of Taiwan through grant no. 111B3005191. L.C. acknowledges support from FONDECYT grant no. 1211656 and the Millennium Nucleus YEMS, NCN2021-080, from ANID, Chile. L.I.C. gratefully acknowledges support from the David and Lucille Packard Foundation and NASA ATP 80NSSC20K0529. K.F. acknowledges support from JSPS KAKENHI grant nos. 20H05847 and 21K13967.

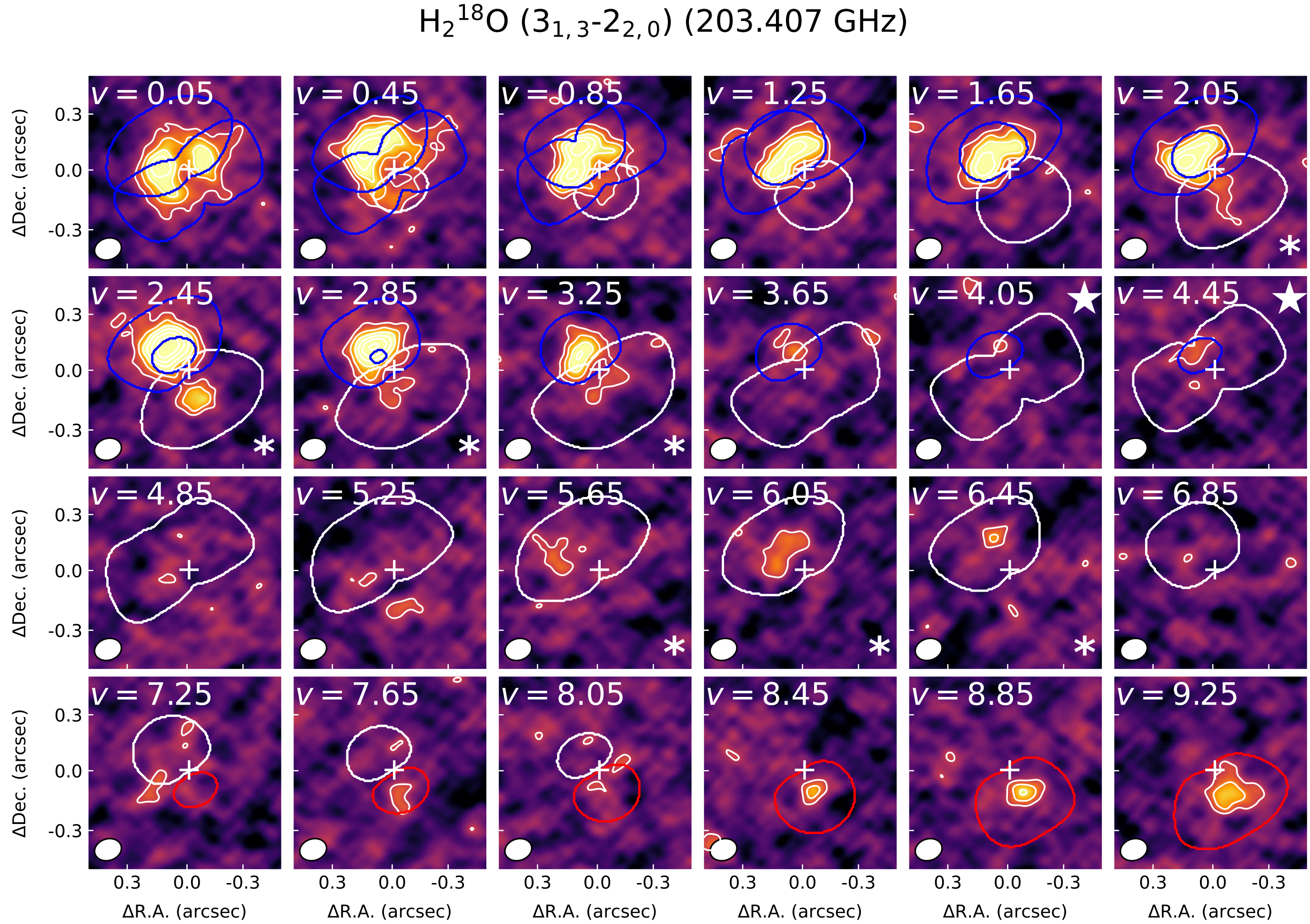
Author contributions J.J.T. wrote the main text and led the data analysis. M.L.R.v.H. assisted with the analysis and writing. M.L. assisted with the snowline analysis and writing T.P.-C. assisted with the snowline analysis and writing. E.F.v.D. and K.F. contributed to the interpretation of results. M.V.P. created two of the figures and contributed to the interpretation of results. L.I.C., D.H., P.D.S. and L.C. contributed to the interpretation of the results and the proofing of the manuscript. All authors contributed to obtaining the observations.

Competing interests The authors declare no competing interests.

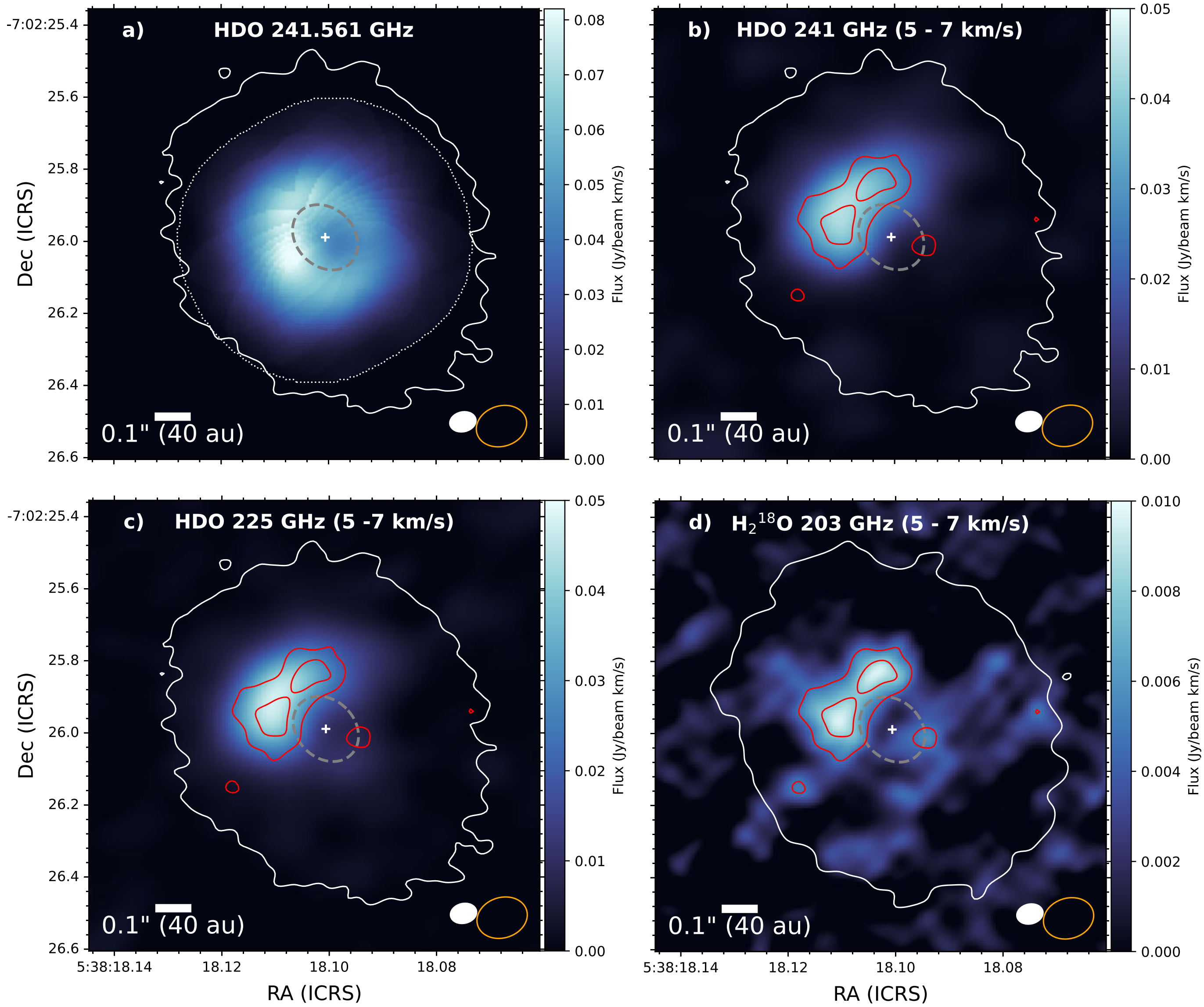
Additional information
Correspondence and requests for materials should be addressed to John J. Tobin.
Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



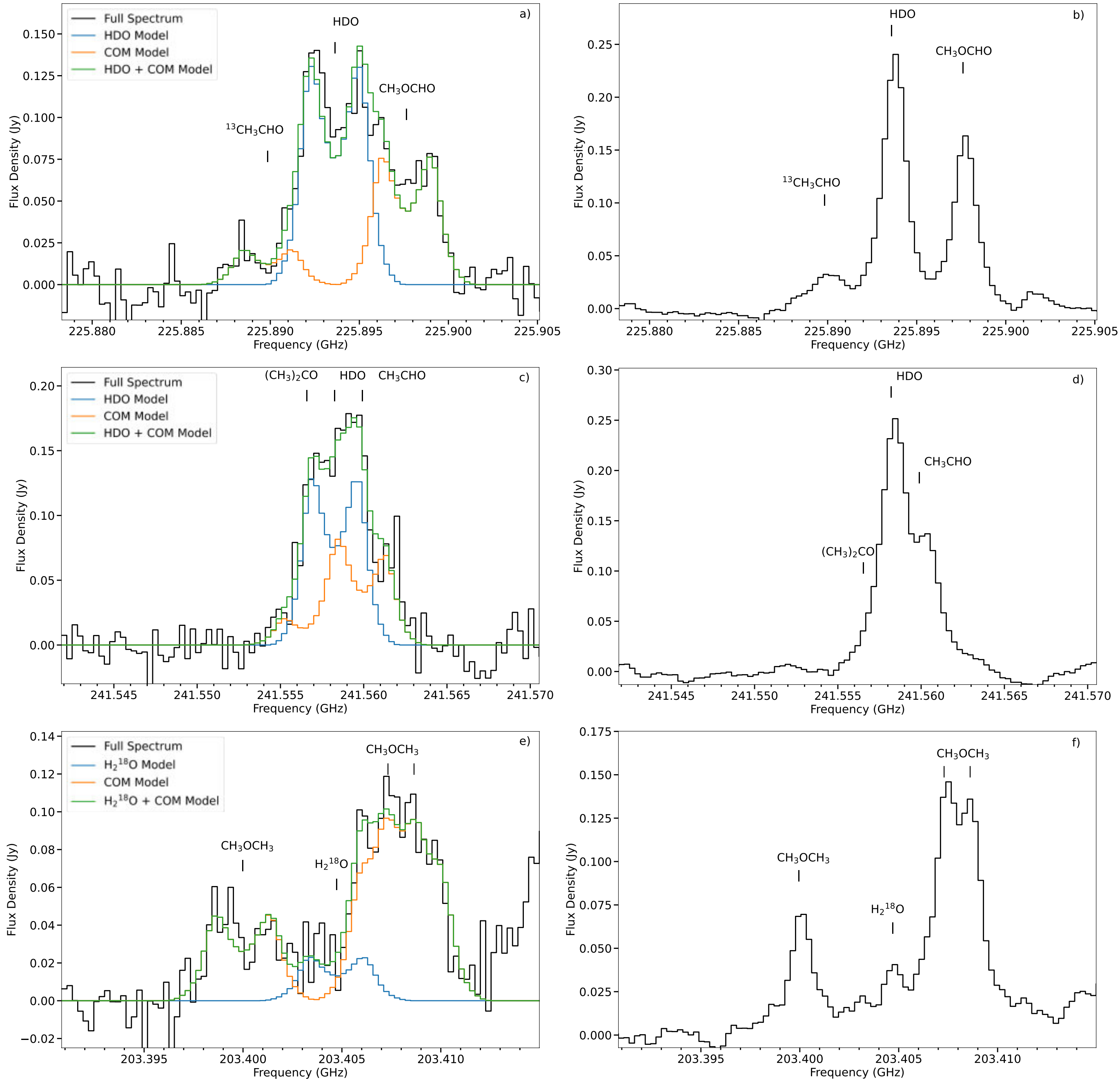
Extended Data Fig. 1 | Channel maps of HDO 225 and 241 GHz emission. The data are shown by the colour scale, the Keplerian mask is drawn as a heavy white line, and the mask corresponding to the blueshifted CH_3OCHO line (Extended Data Fig. 4) is drawn as a heavy blue line for HDO 225 GHz and CH_3CHO for HDO 241 GHz. The continuum peak/protostar position is marked by the white cross. The channels nearest the velocity of V883 Ori (4.25 km s^{-1}) are marked with a star in the upper right corner. The synthesized beam is also drawn in the lower left corner of each panel.



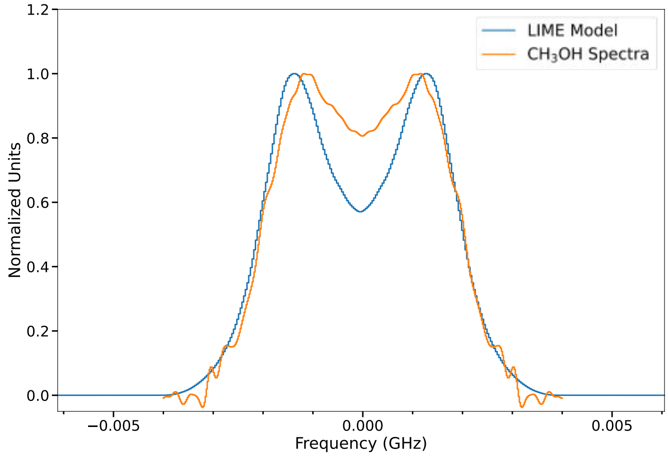
Extended Data Fig. 2 | Channel maps of $H_2{}^{18}O$ 203 GHz emission. The data are shown by the colour scale, the Keplerian mask for $H_2{}^{18}O$ is drawn as a heavy white line, and the masks corresponding to the two blueshifted and one redshifted CH_3OCH_3 lines (Extended Data Fig. 4) are drawn as heavy blue and red lines, respectively. The continuum peak/protostar position is marked by the white cross and the channels nearest the velocity of V883 Ori (4.25 km s^{-1}) are marked with a star in the upper right corner. Despite the faintness of the $H_2{}^{18}O$ line and its location between nearby COM lines, the detection of $H_2{}^{18}O$ is unambiguous given that its emission is detected in the expected channels for the given protostar mass. The channels with >3σ detections within the Keplerian masks are marked with asterisks in the lower right corner. The synthesized beam is drawn in the lower left corner of each panel.



Extended Data Fig. 3 | Integrated intensity images of HDO and $H_2^{18}O$. The HDO 241 GHz integrated intensity map created using a Keplerian mask identical to the HDO 225 GHz and $H_2^{18}O$ lines is shown in **a**, the HDO 241 GHz integrated intensity image between 5.05 and 7.05 km s^{-1} is shown in **b**, the HDO 225 GHz integrated intensity image using the same velocity range is shown in **c**, and the $H_2^{18}O$ integrated intensity image from the same velocity range is shown in **d**. The contours shown in **b**–**d** are from the $H_2^{18}O$ image and show the intensity levels 3 and 5 times the noise (s.d.), where 1 s.d. is 0.00158 Jy beam^{-1} km s^{-1}, and the integrated intensity images in these panels were computed without the use of masks or any other clipping. The 5.05 to 7.05 km s^{-1} velocity range had minimal contamination from other lines for HDO and $H_2^{18}O$ and effectively demonstrates the significance of the $H_2^{18}O$ detection. The extent of the continuum emission from the disk is denoted by the white contour and the position of the protostar is marked with the white cross. The HDO 241 GHz line shows very similar structure to the HDO 225 GHz line that is shown in Fig. 2 in the main text, and the dotted line in **a** shows the region over which the integrated intensity image was computed using the Keplerian mask. The depression in the centre of the line emission in **a** is the result of optically thick continuum absorbing the line emission in the inner ~ 0.1″ (40 au). The extent of this optically thick region is denoted with the thick grey line in the centre of each image. The ellipses in the lower right corner denote the resolution of the line observation (orange, ~ 0.1″) and the continuum data (white, ~ 0.08″).



Extended Data Fig. 4 | Integrated spectra of V883 Ori centered on the HDO 225 GHz, HDO 241 GHz, and $H_2{}^{18}O$ 203 GHz lines. Panels **a**, **c**, **e** show the spectra as observed (disk rotation causes all spectral lines to have a double-peaked line profile), while panels **b**, **d**, **f** show the stacked spectra[40,41] with the Keplerian rotation profile removed. The root mean squared (RMS) noise of the HDO 225 GHz, HDO 241 GHz, and $H_2{}^{18}O$ are 0.016, 0.017, and 0.011 Jy, respectively. The HDO lines are the brightest features around their centre frequencies, but both have contaminating emission from COM species nearby. The $H_2{}^{18}O$ line is faint relative to its surrounding features but is still clearly detected. We are able to model the spectral profiles for HDO, $H_2{}^{18}O$, and the contaminating lines to measure their line fluxes using an optically-thin synthetic spectral model for a disk (Extended Data Fig. 5). In **a**, **c**, **e**, the observed spectrum is drawn as the black line, the model of the contaminating lines is drawn as an orange line, the model HDO and $H_2{}^{18}O$ lines are drawn as blue lines, and the total model of contaminating lines with HDO and $H_2{}^{18}O$ is drawn as a green line. The rise seen toward higher frequencies on the $H_2{}^{18}O$ spectrum (**e**) is another CH_3OCH_3 line that peaks outside the shown region. The spectra are plotted at their observed frequencies and are not corrected for the system local standard of rest (LSR) velocity of ~ 4.25 km s^{-1}.



Extended Data Fig. 5 | Plot of template spectra derived from the LIME radiative transfer model and from the observed isolated methanol lines. The main difference between the templates is at the centre of the line profile where the methanol line has a much more shallow dip owing to line optical depth, while the optically thin LIME model has a much deeper dip at the centre and sharper peaks.

Extended Data Table 1 | Spectral setup

Line	Rest Frequency (GHz)	Band Width (MHz)	Spectral Resolution (KHz, km s^{-1})	Beam (")
Band 5 Data				
$H_2^{18}O$ 3(1,3)- 2(2,0)	203.40752	234.375	122.070, 0.18	0.142×0.112
CH_3CN[a]	202.351612	234.375	122.070, 0.18	0.143×0.113
Continuum	190.5	1875	1129, 1.777	0.150×0.122
Continuum	188.7	1875	1129, 1.794	0.151×0.121
Band 6 Data				
H_2CO 3(1,2)-2(1,1)	225.697775	58.594	122.070, 0.162	0.105×0.083
CH_3OD 5(0,5)-4(0,4) A	226.538674	58.594	122.070, 0.162	0.105×0.083
$C^{17}O$ J=2-1	224.714373	58.594	122.070, 0.162	0.106×0.083
HDO 3(1,2)-2(2,1)	225.89672	58.594	122.070, 0.162	0.104×0.082
CN[a]	226.6595584	117.188	141.113, 0.187	0.102×0.081
CN[a]	226.87478	117.188	141.113, 0.186	0.102×0.081
Continuum	240.5	1875	1129, 1.407	0.099×0.078
HDO 2(1,1)-2(1,2)	241.56155	117.188	122.070, 0.152	0.098×0.077
CH_3OH[a]	241.83165	117.188	122.070, 0.151	0.098×0.078
CH_3OH[b] 5(4)- - 4(4)-	241.806524	"	"	"
CH_3OH[b] 5(-4) - 4(-4) E2	241.813255	"	"	"
CH_3OH[b] 5(-3) - 4(-3) E2	241.852299	"	"	"
CH_3OH[b] 5(1) - 4(1) E1	241.879025	"	"	"
CH_3OH[b] 5(2)+ - 4(2)+	241.887674	"	"	"

Note: while spectral windows are identified as targeting a particular line or continuum, they also generally contain spectral features from other molecules.

[a]Several features of this molecule are contained within the spectral window.

[b]Methanol lines that were isolated enough to use for creation of a template spectrum.

Extended Data Table 2 | Keplerian mask parameters

Datacube	Line	V_{sys} (km s^{-1})	Smoothing (Frac. Beam)	R_{max} (arcsec)	Z/R	ΔV (km s^{-1})
HDO 225 GHz						
	HDO 3(1,2)-2(2,1)	4.25	1.25	0.3	0.4	0.35
	CH_3OCHO 225.90074 GHz	-1.09[b]	1.25	0.3	0.4	0.35
HDO 241 GHz						
	HDO 2(1,1)-2(1,2)	4.25	0.5	0.3	0.2	0.35
	CH_3CHO 241.5632277 GHz	2.17[b]	0.5	0.3	0.2	0.35
H_2CO						
	H_2CO 3(1,2)-2(1,1)	4.25	1.75	0.9	0.2	1.0
$C^{17}O$						
	$C^{17}O$ (J=2-1)[a]	3.7	1.25	0.85	0.2	1.2
	$C^{17}O$ (J=2-1)[a]	4.6	1.25	0.85	0.2	1.2
CH_3OH						
	241.887674 GHz	-50.4[b]	1.75	0.4	0.4	0.35
	241.879025 GHz	-39.7[b]	1.75	0.4	0.4	0.35
	241.852299 GHz	-6.5[b]	1.75	0.4	0.4	0.35
	241.813255 GHz	41.9[b]	1.75	0.4	0.4	0.35
	241.806524 GHz	50.3[b]	1.75	0.4	0.4	0.35
$H_2{}^{18}O$						
	$H_2{}^{18}O$ 3(1,3)- 2(2,0)	4.25	1.25	0.3	0.4	0.35

Note: All masks assumed a stellar mass of 1.285 $M_{\odot}$, an inclination of 38.3°, a position angle of 32°, a distance of 400 pc, and an inner radius of 0.1″. H_2CO and $C^{17}O$ both used an inner radius of 0.05″.

[a]RMS noise of a line-free region of the spectrum.

[b]Offset velocity from the rest frequency of the cube, adopted source velocity was 4.25 km s^{-1}.

Extended Data Table 3 | Spectral lines blended with HDO and $H_2^{18}O$

Molecule	Transition	Rest Frequency (GHz)	Upper Level Energy (K)	Einstein A ($\log_{10}$ (s^{-1}))	RMS[a] (Jy)
$H_2^{18}O$					0.011
$H_2^{18}O$	3(1,3)- 2(2,0)	203.40752	203.7	-5.31770	
CH_3OCH_3	3(3, 1)- 2(2, 1) EE	203.4101126	18.1	-4.26493	
CH_3OCH_3	3(3, 0)- 2(2, 1) AE	203.4114309	18.1	-4.02956	
CH_3OCH_3	3(3, 1)- 2(2, 1) EA	203.4027779	18.1	-4.18698	
HDO 225 GHz					0.016
HDO	3(1,2)-2(2,1)	225.89672	167.6	-4.87987	
CH_3OCHO	6(6,0)- 5(5,0) E	225.90074	36.3	-4.58474	
$^{13}CH_3CHO$[b]	12(2,10) - 11(2,9) E, vt=1	225.8929484	285.7	-3.40579	
HDO 241 GHz					0.017
HDO	2(1,1)-2(1,2)	241.56155	95.2	-4.92559	
CH_3CHO	21(2,19)-20(3,18) E	241.5632277	226.4	-4.70342	
$(CH_3)_2CO$	13(10, 4)-12(9, 3) AA	241.5598869	77.7	-3.48114	

[a]RMS noise of a line-free region of the spectrum.

[b]Tentatively identified line.

Extended Data Table 4 | Impact of different spectral models on line flux

Model	Temp. (K)	Q(T)[a] (HDO, $H_2^{18}O$)	HDO/H_2O	HDO(225) (Jy km s^{-1})	HDO(241) (Jy km s^{-1})	$H_2^{18}O$ (Jy km s^{-1})
Optically-thin	199±42	80.6, 98.8	$(2.26\pm0.63)\times10^{-3}$	0.644±0.028	0.595±0.037	0.126±0.025
Methanol	210±48	86.7, 106.2	$(1.85\pm0.51)\times10^{-3}$	0.64±0.024	0.581±0.04	0.155±0.025
Both[b]	276±87	130.6, 159.8	$(1.92\pm0.49)\times10^{-3}$	0.653±0.026	0.546±0.040	0.159±0.024
Kep. Mask [c]	112±3	35.2, 43.3	$(1.91\pm0.22)\times10^{-3}$	0.554±0.007	0.680±0.007	0.111±0.009
5.05 to 7.05 km s^{-1} [d]						
Kep. Mask[e]	162±10	59.5, 73.0	$(2.54\pm0.44)\times10^{-3}$	0.205±0.004	0.206±0.004	0.034±0.005
Kep. Mask (scaled)[f]	162±10	59.5, 73.0	$(2.54\pm0.44)\times10^{-3}$	0.554±0.007	0.555±0.016	0.093±0.014
Moment 0[g]	182±19	70.7, 86.6	$(2.05\pm0.41)\times10^{-3}$	0.244±0.007	0.233±0.008	0.052±0.008
Moment 0 (scaled)[h]	182±19	70.7, 86.6	$(2.05\pm0.41)\times10^{-3}$	0.658±0.018	0.63±0.020	0.14±0.023

[a]Calculated value of the partition function at the listed temperature.
[b]The optically thin model is used for the HDO and $H_2^{18}O$ lines while the methanol model is used for the COM lines.
[c]Line fluxes measured straight from Keplerian masks, without fine-tuned masks to avoid contamination for $H_2^{18}O$ or HDO 241 GHz. As such, the HDO 241 GHz and $H_2^{18}O$ line fluxes are overestimated and the excitation temperature is underestimated.
[d]Measurements are based line flux measurements only within 5.05 to 7.05 km s^{-1}.
[e]HDO(241) has its flux reduced by 0.007 Jy km s^{-1} to account for minor contamination from CH_3CHO in its Keplerian mask.
[f]The ratio of HDO (225) line flux from 5.05 to 7.05 km s^{-1} to its total line flux (0.37±0.008) is used to scale the HDO(241) and $H_2^{18}O$ line fluxes to their expected values for the full velocity range.
[g]Line fluxes are summed in a 0.4″ radius aperture.
[h]Line fluxes are summed in a 0.4″ radius aperture, and scaled using the ratio of HDO (225) line flux from 5.05 to 7.05 km s^{-1} to its total line flux (0.37±0.008) from the Keplerian Mask measurement.

Extended Data Table 5 | D/H measurements

Object	HDO/H_2O	Method	Reference
Earth			
Earth's Oceans	$(3.11 \pm 0.002)\times 10^{-04}$	VSMOW	[60, 61]
Oort Cloud comets			
1/P Halley	$(4.20 \pm 0.6)\times 10^{-4}$	H_2DO^+/H_3O^+	[62]
C/1996 B2 Hyatuake	$(5.80 \pm 2.0)\times 10^{-4}$	HDO/H_2O	[63]
C/1995 O1 Hale-Bopp	$(6.60 \pm 1.6)\times 10^{-4}$	HDO/H_2O	[64]
C/2007 B3 Lulin	$< 1.12\times 10^{-3}$	HDO/H_2O	[65]
8P/Tuttle	$(8.20 \pm 3.0)\times 10^{-4}$	HDO/H_2O	[66]
C/2009 P1 Garradd	$(4.12 \pm 0.44)\times 10^{-4}$	HDO/H_2O ($H_2{}^{18}O$)	[67]
C/2002 T7 LINEAR	$(5.00 \pm 1.4)\times 10^{-4}$	OD/OH (^{18}OH)	[68]
153P Ikeya-Zhang	$(5.60 \pm 0.6)\times 10^{-4}$	HDO/H_2O ($H_2{}^{18}O$)	[69]
C/2012 F6 Lemmon	$(6.50 \pm 1.6)\times 10^{-4}$	HDO/H_2O ($H_2{}^{18}O$)	[70]
C/2014 Q2 Lovejoy	$(1.40 \pm 0.4)\times 10^{-4}$	HDO/H_2O ($H_2{}^{18}O$)	[70]
Jupiter Family comets			
45P Honda-Mrkos-Pajdusakov (HMP)	$< 4.00\times 10^{-4}$	HDO/H_2O ($H_2{}^{18}O$)	[71]
103P Hartley 2	$(3.20 \pm 0.5)\times 10^{-4}$	HDO/H_2O ($H_2{}^{18}O$)	[72]
67/P Churyumov-Gerasimenko	$(1.06 \pm 0.14)\times 10^{-3}$	HDO/H_2O ($H_2{}^{18}O$)	[24, 26]
46P/Wirtanen	$(3.20 \pm 1.3)\times 10^{-4}$	HDO/H_2O ($H_2{}^{18}O$)	[73]
Flat Spectrum Disk			
V883 Ori	$(2.26 \pm 0.63)\times 10^{-3}$	HDO/H_2O ($H_2{}^{18}O$)	This work
Clustered Class 0 Protostars			
NGC1333 IRAS 4A-NW	$(5.40 \pm 1.5)\times 10^{-4}$	HDO/H_2O ($H_2{}^{18}O$)	[21]
NGC1333 IRAS 2A	$(7.40 \pm 2.1)\times 10^{-4}$	HDO/H_2O ($H_2{}^{18}O$)	[21]
NGC1333 IRAS 4B	$(5.90 \pm 2.6)\times 10^{-4}$	HDO/H_2O ($H_2{}^{18}O$)	[21]
IRAS 16293–2422	$(9.20 \pm 2.6)\times 10^{-4}$	HDO/H_2O ($H_2{}^{18}O$)	[21]
Isolated Class 0 Protostars			
BHR71 IRS	$(1.80 \pm 0.4)\times 10^{-3}$	HDO/H_2O ($H_2{}^{18}O$)	[22]
B335	$(1.70 \pm 0.3)\times 10^{-3}$	HDO/H_2O ($H_2{}^{18}O$)	[22]
L483	$(2.20 \pm 0.4)\times 10^{-3}$	HDO/H_2O ($H_2{}^{18}O$)	[22]

Note: not all measurements use the same method, but when the method is not HDO/H_2O, their values are translated into the equivalent HDO/H_2O by multiplying by a factor of 2 because $HDO/H_2O = 2\times D/H$. See refs. [21,22,24,26,60–73].

Article

Atomic Bose–Einstein condensate in twisted-bilayer optical lattices

https://doi.org/10.1038/s41586-023-05695-4

Received: 11 October 2021

Accepted: 3 January 2023

Published online: 22 February 2023

Check for updates

Zengming Meng[1,5], Liangwei Wang[1,5], Wei Han[1], Fangde Liu[1], Kai Wen[1], Chao Gao[2], Pengjun Wang[1], Cheng Chin[3] & Jing Zhang[1,4] ✉

Observation of strong correlations and superconductivity in twisted-bilayer graphene[1–4] has stimulated tremendous interest in fundamental and applied physics[5–8]. In this system, the superposition of two twisted honeycomb lattices, generating a moiré pattern, is the key to the observed flat electronic bands, slow electron velocity and large density of states[9–12]. Extension of the twisted-bilayer system to new configurations is highly desired, which can provide exciting prospects to investigate twistronics beyond bilayer graphene. Here we demonstrate a quantum simulation of superfluid to Mott insulator transition in twisted-bilayer square lattices based on atomic Bose–Einstein condensates loaded into spin-dependent optical lattices. The lattices are made of two sets of laser beams that independently address atoms in different spin states, which form the synthetic dimension accommodating the two layers. The interlayer coupling is highly controllable by a microwave field, which enables the occurrence of a lowest flat band and new correlated phases in the strong coupling limit. We directly observe the spatial moiré pattern and the momentum diffraction, which confirm the presence of two forms of superfluid and a modified superfluid to insulator transition in twisted-bilayer lattices. Our scheme is generic and can be applied to different lattice geometries and for both boson and fermion systems. This opens up a new direction for exploring moiré physics in ultracold atoms with highly controllable optical lattices.

New band structures in lattice systems often lead to new material functions and discoveries. Twistronics, originating from the twisted-bilayer-graphene as a tuneable experimental platform[1–8], has attracted broad attention in recent years and launched intensive theoretical research. Here, overlaying two graphene layers with a small relative angle show the rich phase diagram, such as the coexistence of unconventional superconductivity and correlated insulating phases[2–4]. In recent years, many examples of twisted-bilayer are discovered with remarkable physical properties not present in their untwisted counterparts. Recently, photonic moiré lattices are explored for their capabilities in localizing and delocalizing light[13–15] and engineering the photonic dispersion of phonon polaritons[16].

Ultracold atoms in optical lattices constitute an ideal platform to simulate emerging many-body phenomena in condensed matter physics[17–19]. Different optical lattice geometries can be realized by interfering different sets of laser beams[20–25]. In particular, a scheme of simulating twisted-bilayer lattice has recently been proposed using two overlapping optical lattices[26,27]. Other schemes for simulating bilayer heterostructures have also been put forward[28,29]. These schemes are based on coherent coupling between spin states of atoms, which simulates interlayer tunnelling along an artificial, synthetic dimension[30–32].

In this article, we demonstrate Bose–Einstein condensates (BEC) of Rubidium-87 (^{87}Rb) atoms loaded into a pair of twisted-bilayer optical lattices. Two overlapping lattices V_1 and V_2 are formed by interfering laser beams at the specific 'tune-out' wavelengths[33–35] λ_1 and λ_2 with proper polarizations such that atoms in spin state $|1\rangle \equiv |F=1, m_F=1\rangle$ and state $|2\rangle \equiv |F=2, m_F=0\rangle$ only experience the lattice potential V_1 and V_2, respectively (Fig. 1). Here F and m_F are the angular momentum and projection quantum numbers in the ^{87}Rb ground state manifold. Each set of the laser beams forms a two-dimensional (2D) square lattice on the horizontal xy plane and the twist of the two lattices is realized by orienting the beams of different wavelengths with a small relative angle $\theta = 5.21°$. The sample is tightly confined in the vertical z direction such that the sample is in the quasi-2D regime (see Methods for details).

The two spin states of ^{87}Rb atoms constitute the synthetic dimension that accommodates the two twisted layers of lattices V_1 and V_2. To precisely determine the tune-out wavelengths λ_1 and λ_2 of the optical lattices V_1 and V_2, we measure the diffraction of atoms by the optical lattices. The experimental sequence starts with an almost pure BEC in a crossed-beam dipole trap. The atoms are prepared in one of the two spin states and a short pulse of the lattice beams is applied. The lattice potential induces Bragg diffraction of atoms to high momentum states. After turning off the lattice beams, we image the diffracted atoms. The wavelengths of the lattice beams are finely adjusted to the tune-out wavelengths such that atoms in state $|1\rangle$ are only diffracted by the lat-

[1]State Key Laboratory of Quantum Optics and Quantum Optics Devices, Institute of Opto-Electronics, Collaborative Innovation Center of Extreme Optics, Shanxi University, Taiyuan, P. R. China. [2]Department of Physics, Zhejiang Normal University, Jinhua, P. R. China. [3]James Franck Institute, Enrico Fermi Institute, Department of Physics, University of Chicago, Chicago, IL, USA. [4]Hefei National Laboratory, Hefei, P. R. China. [5]These authors contributed equally: Zengming Meng, Liangwei Wang. ✉e-mail: jzhang74@sxu.edu.cn

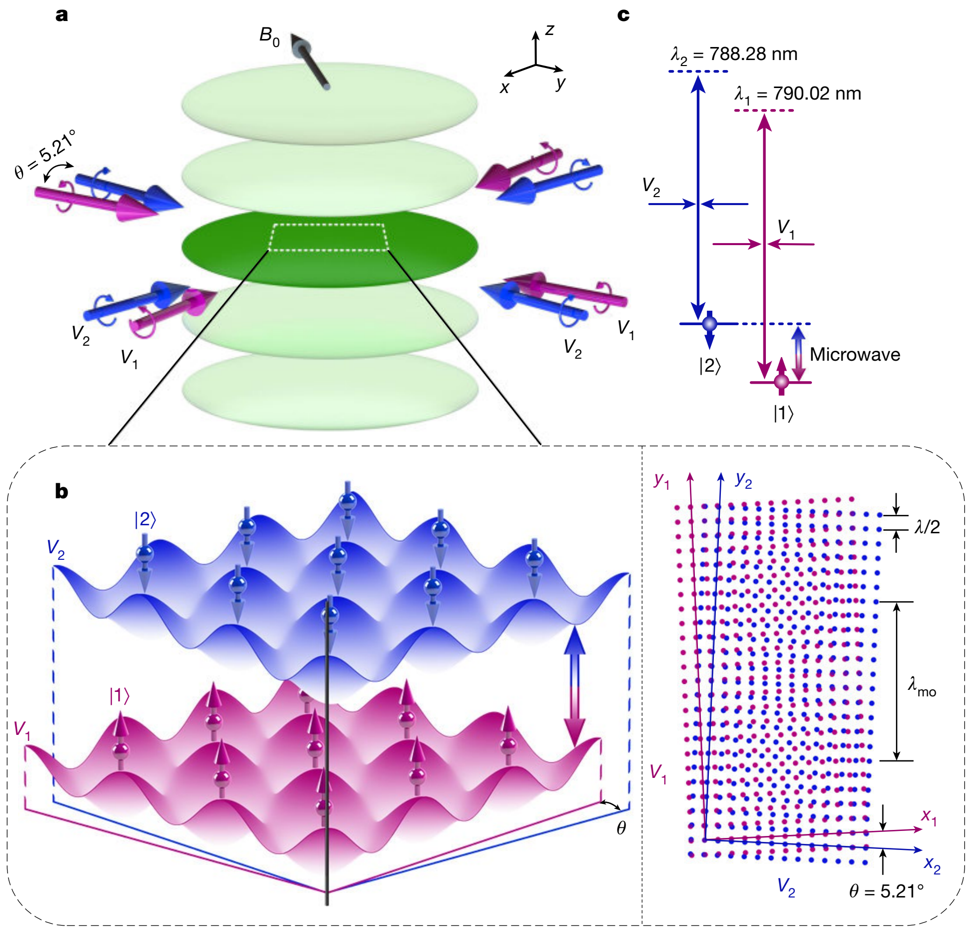


Fig. 1 | Simulation of twisted-bilayer systems based on atoms in spin-dependent optical lattices. a, Atoms are loaded into a single layer, 2D pancake-like potential formed by a vertical optical lattice (green) in the *z* direction. Two sets of square optical lattices V_1 (purple) and V_2 (blue) on the horizontal plane with a small relative angle $\theta = 5.21°$ form a spin-dependent lattice potential and confine Rb atoms in spin state |1⟩ (up arrows) and |2⟩ (down arrows) independently. A magnetic field is applied in the *xy* plane along the 45° diagonal of the V_2 lattice. The lattice beams for V_1 and V_2 are set with opposite circular polarization to generate the vector shift with the opposite sign. **b**, The left panel shows a sketch of the bilayer lattices in the synthetic dimension. The interlayer tunnelling is controlled by a microwave field. The right panel shows a superimposed lattice structure with the lattice constant $\lambda/2$ and much larger moiré length λ_{mo}. **c**, Energy diagram of the two ground Zeeman states |1⟩ and |2⟩ and the associated lattice beams at the tune-out wavelengths $\lambda_1 = 790.02$ and $\lambda_2 = 788.28$ nm.

tice potential V_1 and not by the potential V_2 as shown in Fig. 2. Similarly, atoms in state |2⟩ only experience the potential V_2, but not V_1. By eliminating the cross-talks, we determined the tune-out wavelengths to be $\lambda_1 = 790.02$ and $\lambda_2 = 788.28$ nm. The lattice beams are circularly polarized to produce spatial intensity modulation such that the lattice potentials are attractive to atoms in both spin states (see Methods for details).

Experimentally intralayer hoppings t_1 and t_2 between lattice sites are controlled by the depth of the optical lattices V_1 and V_2; interlayer hopping Ω_R, on the other hand, is independently induced by microwaves (MW) that couple the two spin states. Starting with atoms in state |1⟩ in the dipole trap, for example, the MW spectrum shows a single narrow peak when atoms are driven to state |2⟩. By loading the atoms into the twisted-bilayer optical lattices, the spectrum shows several peaks. The peaks correspond to transitions from atoms in the ground band of lattice V_1, which we label |1, *S*⟩, to different Bloch bands of lattice V_2, which we label |2, *S*⟩, |2, *P*⟩, |2, *D*⟩ and so on (Fig. 3a,b). The peak locations agree with the calculated energies of the *s*, *p* and *d* bands in lattice V_2. The multi-peak structure supports that atoms in different spin states are confined in different lattices. If atoms are loaded into a spin-independent lattice, only a single narrow peak shows up in the spectrum, which belongs to the |1, *S*⟩ to |2, *S*⟩ transition. This is because MW transitions between different Bloch bands are negligible in spin-independent lattices. In the twisted optical lattice, the transitions from the *s* band of state |1⟩ to other bands of state |2⟩ are allowed. In the presence of the twisted-bilayer lattices, the transitions are broadened as the two spin states experience different trap potentials, which induce fast dephasing. Moreover, the on-site interactions increase in deeper lattice potential, resulting in faster decay from high bands to lower bands and thus broader spectral lines. Our observation supports MW as a versatile and powerful tool to induce interlayer hopping between the two twisted layers in the synthetic (spin) dimension.

To quantify the interlayer hopping energy, we measure the time evolution of the population in state |2⟩. We observe a coherent oscillation at detuning $\Delta = -0.9$ kHz, which corresponds to the transition from |1, *S*⟩ to |2, *S*⟩ (Fig. 3c). The interlayer coupling strength can be determined from the oscillation frequencies. In our experiment, the coupling strength is tuneable up to $1E_r$, which exceeds that in typical twisted-bilayer graphene systems. On the other hand, coupling to the *p* band |2, *P*⟩ leads to faster decay probably due to collisional relaxation to the lower *s* band (Fig. 3d). In the following, we will focus on atoms in the twisted-bilayer optical lattices with MW-induced coupling between the *s* bands of the two layers.

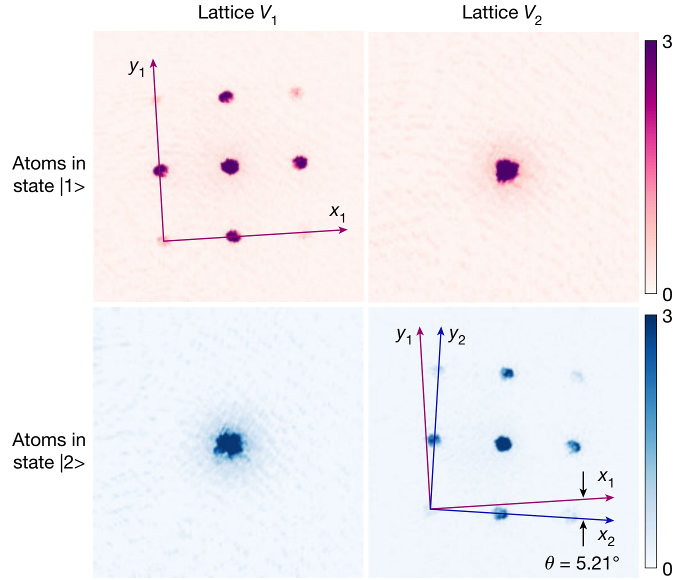


Fig. 2 | Independent diffraction of atoms in different spin states by the twisted-bilayer optical lattices. The optical lattice potential is applied to the atomic BEC with a short duration of 4 µs. The images show diffraction patterns of the atoms after 18 ms of free space expansion. At the tune-out wavelength $\lambda_1 = 790.02$ and $\lambda_2 = 788.28$ nm, atoms in state |1⟩ and |2⟩, are diffracted by the associated optical lattices V_1 and V_2, respectively.

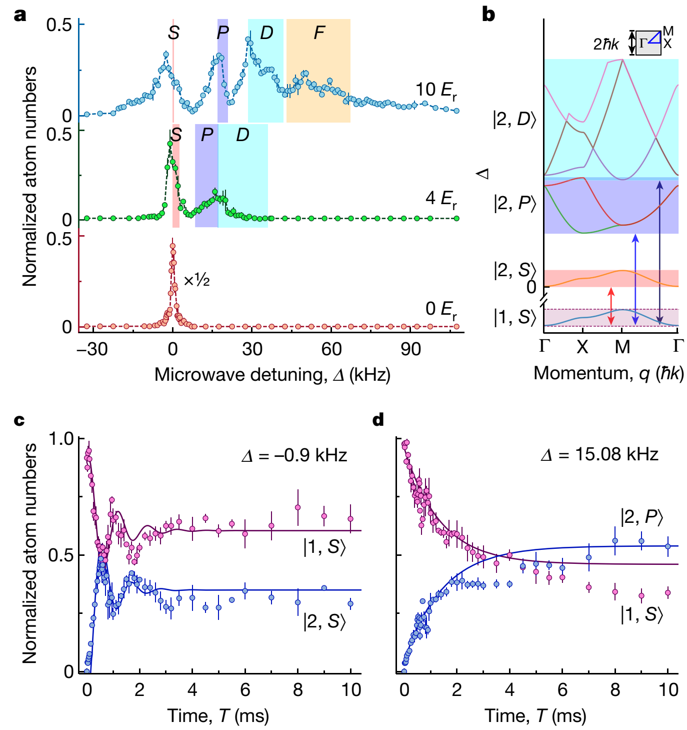


Fig. 3 | Interlayer coupling in twisted-bilayer optical lattices. a, MW spectrum of atoms in the twisted-bilayer optical lattices. Atoms in spin state |1⟩ are driven by MW to spin state |2⟩ in the presence of the lattice potential with depth of 0 (no lattice), 4 and $10E_r$. Here $E_r = q_r^2/2m = h \times 3.67$ kHz is the recoil energy, $q_r = \hbar k = h/\lambda$ is the recoil momentum, m is the atomic mass of ^{87}Rb and λ is the wavelength of the lattice laser. The MW pulse length of 530 µs corresponds to a π pulse in the absence of the lattice potential. **b**, Lattice band structure for the two spin states calculated with the lattice depth $4E_r$. The MW field drives atoms from the s band of state |1⟩, labelled as |1, S⟩ to s, p and d bands of state |2⟩ with different detuning Δ. **c**,**d**, Starting with all atoms in |1, S⟩, population in state |2⟩ is measured in the twisted-bilayer lattices at $4E_r$ after the MW pulse that drives the atoms to |2, S⟩ with the detuning $\Delta = -0.9$ kHz (**c**), or to |2, P⟩ with detuning $\Delta = 15.08$ kHz (**d**). Fits in **c** show an interlayer coupling frequency of $\Omega_R = 2\pi \times 893$ Hz and a decay rate of 1,200 s^{-1}. Lines in **d** are guides to the eye. Each point is based on three or more measurements and error bars show the standard deviations of the mean.

A key signature of atoms in the twisted-bilayer optical lattice is the moiré lattice with a period

$$\lambda_{mo} = \frac{a}{2\sin\theta/2}, \quad (1)$$

which, for the lattice constant $a = 395$ nm and twist angle $\theta = 5.21°$, amounts to a moiré length $\lambda_{mo} = 4.35$ µm. The large moiré period gives rise to a mini-Brillouin zone in the momentum space, which is expected to generate the flat bands and strongly correlated states[1–12]. Notably, an in situ moiré pattern is also observed in one-dimensional lattices with two lattice constants[36]. To identify the moiré length scale in our system, we use in situ absorption imaging to visualize the moiré pattern (Fig. 4a–f). Here we first load the atoms in state |1⟩ into the lowest s band of lattice V_1 and then ramp up the MW field with detuning $\Delta = -0.9$ kHz to drive the transition from |1, S⟩ to |2, S⟩. We then in situ image the atoms in state |2⟩. Moiré patterns in one and two dimensions are observed, and the moiré period is measured to be 4.35 µm consistent with expectation (Fig. 4a–f). Note that the primary optical lattice spacing $a = 395$ nm is indiscernible with our imaging optics.

We also examine the quantum state of atoms in the bilayer twisted lattices by analysing their momentum-space distribution. After loading a BEC into the bilayer lattice of $4E_r$ in the presence of resonant MW transition, we hold for some time and then perform the time-of-flight (TOF) measurement (Fig. 4g,h). Two sets of diffractions manifest, which correspond to the primary lattice momentum π/a and the much smaller moiré momentum π/λ_{mo}. The high contrast of both sets of diffraction pattern suggests that the atoms remain in the superfluid phase with phase coherence extending beyond the moiré length scale. In particular, the contrasts of the moiré pattern in real and momentum space persist over 40 ms (Fig. 4i), from which we conclude that the atoms maintain in the superfluid phase in the twisted-bilayer lattices.

Theoretically, depending on the twist angle θ, the superimposed twisted-bilayer lattice can yield either a periodic potential with supercells that supports a delocalized ground state or a quasi-periodic one that supports a localized ground state in the absence of interactions. In fact, only specific twist angles give rise to periodic lattice potentials. For square lattices, the twist angles that lead to commensurate superlattice should satisfy $\theta = 2\arctan(\overline{m}/\overline{n})$, where $\overline{m}$ and $\overline{n}$ are coprime natural number[13]. The twist angle $\theta = 5.21°$ used in our work is close to the commensurate angle $\theta = 2\arctan(1/22) \approx 5.205°$, and the period of the supercell is given by $2\lambda_{mo} \approx 22a$ (see Methods for details). Whereas our twist angle does not exactly match the commensurate angle $\theta = 2\arctan(1/22) \approx 5.205°$, the small difference cannot be distinguished in a finite size sample due to repulsive interactions (see Methods for details). In particular, the spatial moiré period remains a clear observable in our experiment because of the finite chemical potential of our atomic superfluid. The persistence of the spatial and momentum periodicity of the sample in the twisted-bilayer lattice supports the superfluid as the ground state of the system.

Compared with electronic materials, in which the flat band is investigated frequently near the Fermi surface, we can also explore flat-band physics with bosons condensed in the lowest band. In our system, when interlayer coupling increases, the long-wavelength moiré potential becomes deeper, so atoms in the lowest band are isolated at a larger spatial scale (moiré wavelength), which flattens the ground band and

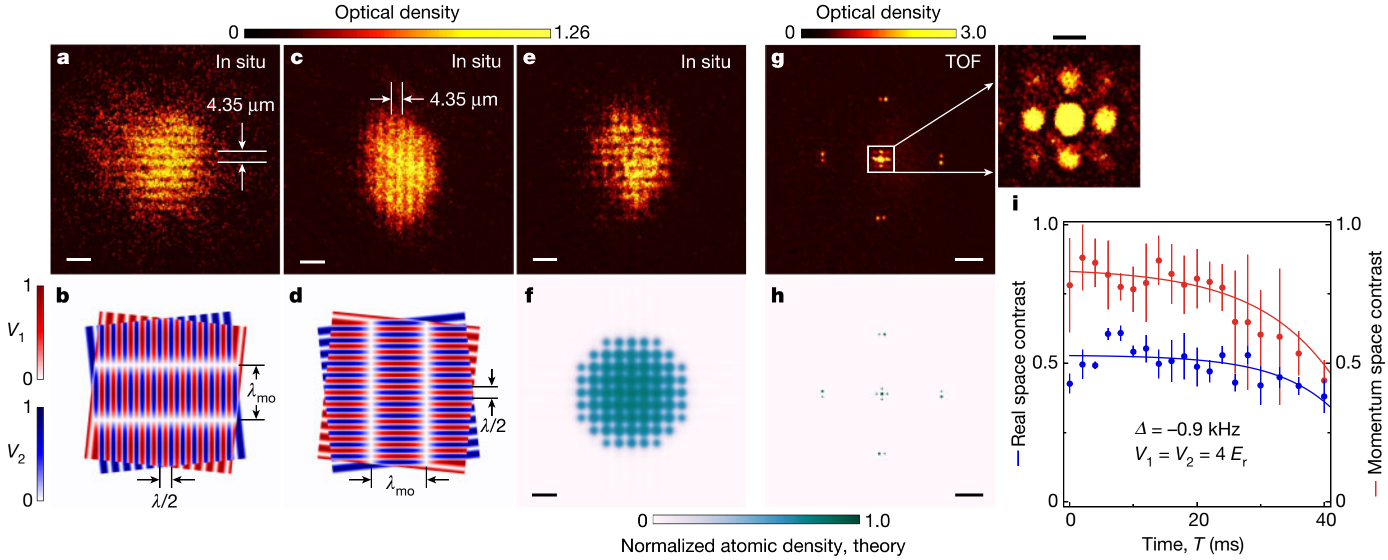


Fig. 4 | Moiré pattern and superfluid ground state in twisted-bilayer optical lattices. a,**c**, Moiré pattern of atoms in one-dimensional bilayer optical lattices in the x (**a**) and y directions (**c**). **b**,**d**, Plot of the corresponding optical lattice potential in the x (**b**) and y directions (**d**). Note that the primary optical lattice $\lambda/2$ is indiscernible in the in situ images. In all experiments, an MW field is applied to couple the s band of state $|1\rangle$ and s band of state $|2\rangle$. The atoms in state $|2\rangle$ are measured. The elliptical shape of the atomic cloud is induced by a little asymmetric harmonic trap potential in xy plane. **e**,**f**, Experimental (**e**) and theoretical (**f**) moiré pattern of atoms in the presence of optical lattices in both directions. **g**,**h**, TOF images with 18 ms from the experiment (**g**) and calculation (**h**) show diffraction peaks associated with the primary lattice constant $\lambda/2$ and moiré length λ_{mo}. The enlarged picture of the centre part of **g** is obtained by the imaging system with higher magnification. **i**, The contrast of moiré pattern in real space and the contrast of diffraction pattern for different hold times. Here, the lattice depths are $V_{1x} = V_{2x} = V_{1y} = V_{2y} = 4E_r$ ($U/t = 1.67$, $\Omega_R/t = 2.07$, U is the on-site interaction). The real-space and momentum-space distributions of **f** and **h** are theoretically calculated by solving the mean-field ground states according to the Gross–Pitaevskii equations (see Methods for details). The contrast in the real space is defined as $(S_{max} - S_{min})/(S_{max} + S_{min})$, where S_{max} and S_{min} are the maximum and minimum atomic density of the moiré fringes. The contrast in the momentum space is defined as $(P^{av}_{max} - P_{min})/(P^{av}_{max} + P_{min})$, where, $P^{av}_{max} = (P^{0}_{max} + P^{1}_{max})/2$ and P_{min} are the average maximum and minimum density of the diffraction pattern. Here P^{0}_{max} is zero-momentum component and P^{1}_{max} is the moiré component near the zero momentum. Error bars show the standard deviation of the mean. Scale bars **a**,**c**,**e**,**f**, 10 μm, **g**,**h**, 100 μm and inset of **g**, 20 μm.

enhances the localization of the atoms. In the large interlayer coupling limit, the system can be regarded as a single layer (single-component) experiencing a twisted optical lattice (see Methods for details). The single-layer system with a twisted optical lattice admits a flat-band structure in the ground band, which has also been studied experimentally in photonic systems[13–15]. The easily tuned intra- and interlayer couplings in our system offer the added advantages of seeking new quantum phases and phase transitions with cold atoms.

By varying the depth of optical lattices and interlayer coupling, we find several distinct quantum phases, including superfluid (SF), superfluid with only short-range coherence (SF-II), Mott insulator (MI) and insulator (I) (Fig. 5a and Methods). These phases can be distinguished

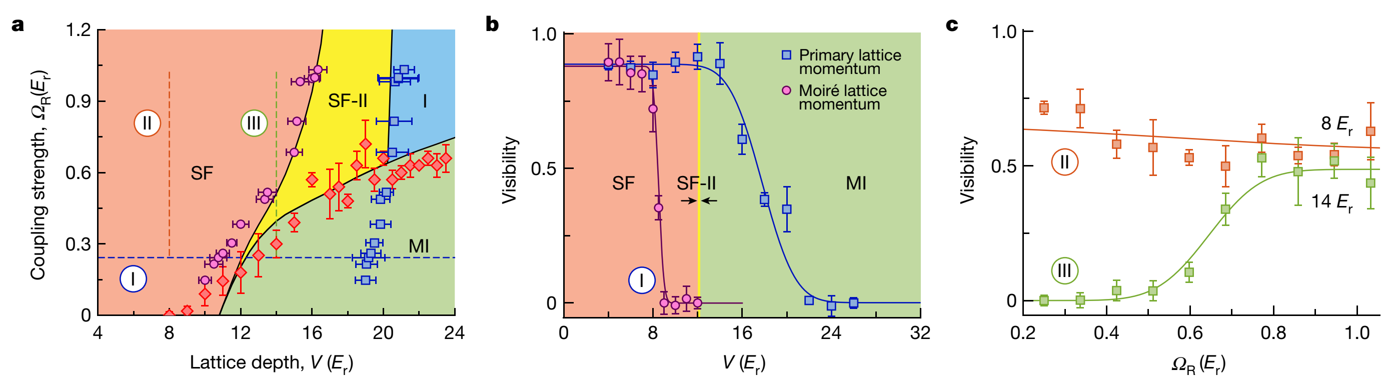


Fig. 5 | Phase transition for the twisted-bilayer optical lattice. a, Phase diagram (see Methods for details), in which SF, SF-II, MI and I refer to superfluid, superfluid only with short-range coherence, Mott insulator and insulator. The solid curves denote the calculated phase boundaries with mean-field at zero temperature. The dots are experimental measurements of the phase boundaries. The pink circles and blue squares denote the loss of the coherence at the moiré length scale and the length scale of the primary lattice, respectively. The red diamonds denote the appearance of the moiré pattern in the real-space in situ images that increases the interlayer coupling. The error bars indicate the experimental uncertainties in determining the phase transition. **b**, Visibility curves for the moiré and primary lattice momentum components as a function of lattice depth (path I in **a**). A sequential loss of phase coherence appears at the moiré momentum and the primary lattice momentum. The intermediate regime indicates SF-II phase. The interlayer coupling frequency is $\Omega_R = 0.24E_r$. **c**, Visibility curves for the moiré momentum component as a function of interlayer coupling with the lattice depth $8E_r$ (path II in **a**) and $14E_r$ (path III in **a**), respectively. The solid lines in **b** and **c** are fitted from the experimental data and only guide the eye.

by the phase coherence and real-space density correlations. The SF-II phase emerges with finite interlayer coupling around the transition from a regular SF to an insulator. The spatial range of phase coherence is the key to distinguishing the two SF phases: while an SF supports long-range phase coherence[37], the SF-II phase maintains the coherence only up to the moiré length scale. In addition, the SF-II phase supports the moiré pattern in the real space. Theoretically, SF-II is the phase with superfluid domains embedded in a gapped insulator, induced by the interlayer coupling. Finally, the insulator phases I and MI can be identified by the disappearance of spatial coherence at all scales and integer fillings of all the sites. Whereas the MI has uniform atom density with weak interlayer coupling, the I phase features a moiré pattern due to stronger interlayer coupling.

In the experiment, we measure the phase coherence from the momentum-space diffraction peaks in the TOF images and directly probe the moiré pattern by in situ imaging following the measurement method as shown in Fig. 4. The measurement of the phase boundaries is shown in Fig. 5a. We use three independent paths to study these phases. In path I, we fix the interlayer coupling strength at a small value $\Omega_R = 0.24E_r$ and increase the lattice depth. The phase transition from SF to MI and across SF-II is shown in the TOF images (Fig. 5b). Here the diffraction peaks at the moiré momenta disappear first before the disappearance of the primary lattice. The intermediate regime indicates the SF-II phase in which the moiré-scale long-range correlation is destroyed while a short-range coherence remains; at the same time, the density correlations of moiré pattern appear in the real space. In path II, we fix the lattice depth in the SF region and increase the interlayer coupling. The diffraction peaks at the moiré momenta persist with high contrasts. However, in path III, when the depth of optical lattices is fixed at the MI region and the interlayer coupling increases, the visibility at the moiré momenta presents the threshold behaviour and emerges at $\Omega_R > 0.5E_r$ (Fig. 5c). These observations are qualitatively consistent with the theoretical expectation and demonstrate that the interlayer coupling can induce a re-entrant transition from MI to SF across SF-II. One may understand such rich transitions from the fact that the interplay between the interlayer coupling and interactions tends to localize the bosons, primarily in the moiré length scale.

This work provides a preliminary physical insight into the quantum phase transition between SF and SF-II (MI and SF-II or MI and I) and offers the possibility to study the complex phases due to the presence of quasi-disorder induced by large interlayer coupling and strong interaction, such as Bose glass insulator, resembling that in disordered bosonic systems[38–40]. These complex phases are worth further investigating in the future.

The present work focuses on the realization and the ground state properties of atoms in the twisted-bilayer optical square lattice. Our success in loading a superfluid into the bilayer lattice demonstrates a new versatile platform to explore moiré physics and the associated superfluidity in a quantum many-body system. Beyond the tuneable twist angle, the cold atom platform offers remarkable controls such as different lattice depths and interlayer coupling in different layers.

Furthermore, the twisted-bilayer square lattice closely connects to the physics of heterostructures of twisted atomically thin semiconductors[8,41,42]. At the same time, our experiment can in principle be extended to multi-layer lattice in which the interlayer couplings can be independently induced by MW and radio-frequencies. Replacing the MW with optical Raman transitions, the interlayer coupling can be spatial dependence, which can support topological ground states. Finally, our optical lattice scheme can be applied to confine fermionic atoms in bilayer hexagonal lattice, which faithfully simulates electrons in a bilayer graphene, and may offer insight into the emergence of superconductivity in the strongly correlated, flat-band regime.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-05695-4.

1. Cao, Y. et al. Correlated insulator behaviour at half-filling in magic-angle graphene superlattices. *Nature* **556**, 80–84 (2018).
2. Cao, Y. et al. Unconventional superconductivity in magic-angle graphene superlattices. *Nature* **556**, 43–50 (2018).
3. Yankowitz, M. et al. Tuning superconductivity in twisted bilayer graphene. *Science* **363**, 1059–1064 (2019).
4. Lu, X. et al. Superconductors, orbital magnets and correlated states in magic-angle bilayer graphene. *Nature* **574**, 653–657 (2019).
5. Wang, J., Mu, X., Wang, L. & Sun, M. Properties and applications of new superlattice: twisted bilayer graphene. *Mater. Today Phys.* **9**, 100099 (2019).
6. Andrei, E. Y. & MacDonald, A. H. Graphene bilayers with a twist. *Nat. Mater.* **19**, 1265–1275 (2020).
7. Balents, L., Dean, C. R., Efetov, D. K. & Young, A. F. Superconductivity and strong correlations in Moiré flat bands. *Nat. Phys.* **16**, 725–733 (2020).
8. Kennes, D. M. et al. Moiré heterostructures as a condensed-matter quantum simulator. *Nat. Phys.* **17**, 155–163 (2021).
9. Lopes dos Santos, J. M. B., Peres, N. M. R. & Castro Neto, A. H. Graphene bilayer with a twist: electronic structure. *Phys. Rev. Lett.* **99**, 256802 (2007).
10. Mele, E. J. Commensuration and interlayer coherence in twisted bilayer graphene. *Phys. Rev. B* **81**, 161405 (2010).
11. Bistritzer, R. & MacDonald, A. H. Moiré bands in twisted double-layer graphene. *Proc. Natl Acad. Sci. USA* **108**, 12233–12237 (2011).
12. Moon, P. & Koshino, M. Energy spectrum and quantum Hall effect in twisted bilayer graphene. *Phys. Rev. B* **85**, 195458 (2012).
13. Wang, P. et al. Localization and delocalization of light in photonic moiré lattices. *Nature* **577**, 42–46 (2020).
14. Huang, C. et al. Localization-delocalization wavepacket transition in Pythagorean aperiodic potentials. *Sci. Rep.* **6**, 32546 (2016).
15. Fu, Q. et al. Optical soliton formation controlled by angle twisting in photonic Moiré lattices. *Nat. Photon.* **14**, 663–668 (2020).
16. Hu, G. et al. Topological polaritons and photonic magic angles in twisted α-MoO_3 bilayers. *Nature* **582**, 209–213 (2020).
17. Bloch, I., Dalibard, J. & Zwerger, W. Many-body physics with ultracold gases. *Rev. Mod. Phys.* **80**, 885–964 (2008).
18. Lewenstein, M., Sanpera, A. & Ahufinger, V. *Ultracold Atoms in Optical Lattices: Simulating Quantum Many-Body systems* (Oxford Univ. Press, 2012).
19. Windpassinger, P. & Sengstock, K. Engineering novel optical lattices. *Rep. Prog. Phys.* **76**, 086401 (2013).
20. Soltan-Panahi, P. et al. Multi-component quantum gases in spin-dependent hexagonal lattices. *Nat. Phys.* **7**, 434–440 (2011).
21. Wirth, G., Ölschläger, M. & Hemmerich, A. Evidence for orbital superfluidity in the P-band of a bipartite optical square lattice. *Nat. Phys.* **7**, 147–153 (2011).
22. Tarruell, L., Greif, D., Uehlinger, T., Jotzu, G. & Esslinger, T. Creating, moving and merging Dirac points with a Fermi gas in a tunable honeycomb lattice. *Nature* **483**, 302–305 (2012).
23. Jo, G. B. et al. Ultracold atoms in a tunable optical kagome lattice. *Phys. Rev. Lett.* **108**, 045305 (2012).
24. Taie, S. et al. Coherent driving and freezing of bosonic matter wave in an optical Lieb lattice. *Sci. Adv.* **1**, e1500854 (2015).
25. Gall, M., Wurz, N., Samland, J., Chan, C. F. & Köhl, M. Competing magnetic orders in a bilayer Hubbard model with ultracold atoms. *Nature* **589**, 40 (2021).
26. González-Tudela, A. & Cirac, J. I. Cold atoms in twisted-bilayer optical potentials. *Phys. Rev. A* **100**, 053604 (2019).
27. Luo, X. & Zhang, C. Spin-twisted optical lattices: tunable flat bands and Larkin-Ovchinnikov superfluids. *Phys. Rev. Lett.* **126**, 103201 (2021).
28. Graß, T., Chhajlany, R. W., Tarruell, L., Pellegrini, V. & Lewenstein, M. Proximity effects in cold atom artificial graphene. *2D Mater.* **4**, 015039 (2016).
29. Salamon, T. et al. Simulating twistronics without a twist. *Phys. Rev. Lett.* **125**, 030504 (2020).
30. Boada, O., Celi, A., Latorre, J. I. & Lewenstein, M. Quantum simulation of an extra dimension. *Phys. Rev. Lett.* **108**, 133001 (2012).
31. Celi, A. et al. Synthetic gauge fields in synthetic dimensions. *Phys. Rev. Lett.* **112**, 043001 (2014).
32. Ozawa, T. & Price, H. M. Topological quantum matter in synthetic dimensions. *Nat. Rev. Phys.* **1**, 349–357 (2019).
33. LeBlanc, L. J. & Thywissen, J. H. Species-specific optical lattices. *Phys. Rev. A* **75**, 053612 (2007).
34. Arora, B., Safronova, M. S. & Clark, C. W. Tune-out wavelengths of alkali-metal atoms and their applications. *Phys. Rev. A* **84**, 043401 (2011).
35. Wen, K. et al. Experimental study of tune-out wavelengths for spin-dependent optical lattice in ^{87}Rb Bose-Einstein condensation. *J. Opt. Soc. Am. B* **38**, 3269 (2021).
36. McDonald, M., Trisnadi, J., Yao, K. & Chin, C. Superresolution microscopy of cold atoms in an optical lattice. *Phys. Rev. X* **9**, 021001 (2019).
37. Gerbier, F. et al. Interference pattern and visibility of a Mott insulator. *Phys. Rev. A* **72**, 053606 (2005).

38. Fisher, M. P. A., Weichman, P. B., Grinstein, G. & Fisher, D. S. Boson localization and the superfluid-insulator transition. *Phys. Rev. B* **40**, 546–570 (1989).
39. Krauth, W., Trivedi, N. & Ceperley, D. Superfluid-insulator transition in disordered boson systems. *Phys. Rev. Lett.* **67**, 2307–2310 (1991).
40. Freericks, J. K. & Monien, H. Strong-coupling expansions for the pure and disordered Bose-Hubbard model. *Phys. Rev. B* **53**, 2691–2700 (1996).
41. Kariyado, T. & Vishwanath, A. Flat band in twisted bilayer Bravais lattices. *Phys. Rev. Res.* **1**, 033076 (2019).
42. Kennes, D. M., Xian, L., Claassen, M. & Rubio, A. One-dimensional flat bands in twisted bilayer germanium selenide. *Nat. Commun.* **11**, 31124 (2020).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2023

Methods

Experimental setup

In our experiment, the ultracold ^{87}Rb atoms in the $|F=2, m_F=2\rangle$ state are prepared in the crossed optical dipole trap[43]. Forced evaporation in the optical trap creates the BEC with up to 5×10^5 atoms. The atoms can be transferred to the $|F=1, m_F=1\rangle$ state through a rapid adiabatic passage induced by MW transition. To load the atoms into the 2D trap, a 532 nm laser beam is deflected by an acousto-optic deflector and then split into two beams with variable spacing adjusted by the acousto-optic deflector. The two beams are focused onto the atoms with a 150 mm aspherical lens. These beams interfere to form a standing wave in the vertical direction with variable separation (accordion lattice). This separation can be varied from 12 down to 3 µm. The advantage of variable spacing is that we can load a three-dimensional (3D) shaped cloud into a single layer of the 2D pancakes at maximum separation and then compress the pancake adiabatically to reach a deep 2D regime. The maximum vertical confinement can reach more than 20 kHz and we optimize it at 1 kHz to observe moiré pattern and superfluid of ultracold atoms.

The twisted-bilayer optical lattices are created by two sets of 2D square lattice V_1 and V_2. A twisted angle of $\theta = 5.21°$ is set between the two lattice potentials, namely, $V_2(r) = V_1(Sr)$, $S=\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}$. The optical lattices V_1 and V_2 are derived from two continuous-wave Ti:sapphire single frequency lasers (M Squared lasers SolsTiS and Coherent MBR-110), respectively. Two lattice beams V_{1x} and V_{1y} of V_1 are frequency-shifted +80 and +95 MHz by two single-pass acousto-optic modulators, respectively. The same applies to the two lattice beams V_{2x} and V_{2y} of lattice V_2. The four lattice beams are coupled into polarization-maintaining single-mode fibres to improve the stability of the beam pointing and achieve better beam-profile quality. After the fibres, each lattice beam is focused by a lens and retroreflected by a concave mirror. To generate the vector light shift, we use the same circular polarization for two lattice beams to produce spatial intensity modulation. In the experiment, we can determine and calibrate this angle by measuring the intersection angle between two lasers and the moiré period from the in situ images. The estimated uncertainty of the two methods is about 0.05°.

We use the MW field to couple the two spin states for manipulating the interlayer coupling. The 6.8 GHz MW signal is amplified by a 10 W solid state amplifier (Kuhne Electronic, KU PA 640720-10A). We place a circulator on the output of the amplifier to reduce reflected power coming back to the amplifier. The MW is emitted out to the atoms by a sawed-off waveguide, which is placed outside the high vacuum glass cell. We use MW cables to transfer MW from the amplifier to the waveguide. With this MW power amplifier, we can reach the maximum interlayer coupling strength of about $1.0E_r$. It is feasible to increase the interlayer coupling strength to about several E_r by using an available higher power amplifier.

Our image system consists of an objective with a numerical aperture of 0.69, working distance of 11 mm and effective focal length of 18 mm. A 900 mm lens after the objective leads to a magnification of ×50 for in situ imaging with an EMCCD (Andor iXon Ultra 897). We also use a 200 mm (400 mm) lens after the objective leads to a magnification of ×11 (22) for the TOF absorption imaging with 18 ms. The atoms are detected by state-selective absorptive imaging. As we choose two ground hyperfine Zeeman states of ^{87}Rb $|F=2, m_F=0\rangle$ of the $F=2$, and $|F=1, m_F=1\rangle$ of the $F=1$ hyperfine manifold as the two internal spin states, we can fully resolve the population in each individual state. For $|F=2, m_F=0\rangle$ state, a 50 µs long imaging pulse of resonant light on the $F=2\to F'=3$ D_2 cycling transition is used to detect the $|2\rangle$ atoms. To detect the $|F=1, m_F=1\rangle$ state, a resonant light pulse on the $F=2\to F'=3$ cycling transition is first used to remove the $|2\rangle$ atoms and then a 50 µs long imaging pulse of resonant light on the $F=2\to F'=3$ is applied at the same time with a repump light (resonant light $F=1\to F'=2$) to detect the $|1\rangle$ atoms.

When studying the superfluid to MI transition, we use the standard method of interference pattern contrast (visibility) to show this transition[37,44]. We first load the atoms in state $|1\rangle$ into the lowest s band of lattice V_1 by ramping up V_1 and V_2 simultaneously with 30 ms, and then ramp up the MW field with 10 ms to drive the transition from $|1, S\rangle$ to $|2, S\rangle$. The atoms are detected by state-selective absorptive imaging with TOF of 18 ms after switching off all lattices and trapping light. In experiment, we first check that BEC is still kept (Extended Data Fig. 1) as ramp up the lattice V_1 (or V_2) to the higher lattice depth than $24E_r$ and then ramp down again, which makes sure to perform the phase transition from SF to MI successfully for the lattice V1 (or V2). When adding the interlayer coupling between two spin states, and at the same time a quasi-disorder is introduced, there are two more mechanics to make the system not completely reversible. One is the finite coherent time between two spin states. When the system is prepared initially in the spin down, the system will become the spin mixture after the interlayer coupling is ramped back down. We define this process as irreversibility. The other is that adiabaticity is broken down by a quasi-period or disordered lattice, which induces not to completely remain in the zero-momentum state after ramping the lattices back down.

Tune-out wavelength for twisted-bilayer optical lattices

The a.c. Stark shift, or light shift, is a light-induced change of energy level. For alkali-metal atoms, the total a.c. Stark shift can be expressed in the irreducible components (including scalar, vector and tensor components) of the polarizability[45]:

$$\begin{aligned}\Delta U &= \Delta U(F, m_F; \omega)\\ &= -I\Bigg[\alpha^{(0)}(\omega) + \alpha^{(1)}(\omega)(\xi \mathbf{e}_k\cdot\mathbf{e}_B)\frac{m_F}{F}\\ &\quad + \alpha^{(2)}(\omega)\frac{3\cos^2\phi - 1}{2}\frac{3m_F^2 - F(F+1)}{F(2F-1)}\Bigg],\end{aligned} \tag{2}$$

where F is the total atomic angular momentum, m_F is the magnetic quantum number, ω is the laser frequency, I is the laser field intensity, ξ is light ellipticity, $\mathbf{e}_k$ and $\mathbf{e}_B$ are unit vectors along the light wave vector and magnetic field quantization axis, respectively, and ϕ is the intersection angle between the linearly polarized component of light field and $\mathbf{e}_B$. This formula comes from the perturbation expansion. Note that the range of values of light ellipticity is $\xi\in[-1, 1]$, $\xi = \pm1$ denotes left and right circular polarization. $\alpha^{(0)}(\omega)$, $\alpha^{(1)}(\omega)$, $\alpha^{(2)}(\omega)$ are the scalar, vector and tensor polarizability, respectively. Scalar shift is spin independent. Vector shift acts like an effective magnetic field to generate the linear Zeeman splitting (light shift proportional to m_F), which depends on the ellipticity of the light and the intersection angle between the laser beam wave vector and magnetic field quantization axis $\mathbf{e}_B$. So, there are two methods to control the vector shift: rotating bias magnetic field and changing light polarization. The tensor part is derived from the linearly polarized light and acts as an effective d.c. electric field.

For the first excited state of alkali-metal atoms, the fine structure interaction induces the spectral lines of the D1 ($5^2S_{1/2}\to 5^2P_{1/2}$) and D2 ($5^2S_{1/2}\to 5^2P_{3/2}$) lines. The coefficients of the scalar, vector and tensor shifts of the ground states $5^2S_{1/2}$ of ^{87}Rb atoms in equation (2) are given by

$$\begin{aligned}\alpha^{(0)}(\omega) &\approx -\frac{\pi c^2\Gamma_{D2}}{2\omega_0^3}\left(\frac{2}{\delta_{D2}} + \frac{1}{\delta_{D1}}\right),\\ \alpha^{(1)}(\omega) &\approx -\frac{\pi c^2\Gamma_{D2}}{2\omega_0^3}\left(\frac{1}{\delta_{D2}} - \frac{1}{\delta_{D1}}\right)g_F F,\\ \alpha^{(2)}(\omega) &\approx 0,\end{aligned} \tag{3}$$

where Γ_{D_2} is the decay rate of the excited state for D_2 line, $\omega_0 = \frac{1}{3}\omega_{D1} + \frac{2}{3}\omega_{D2}$ is the effective frequency, $\delta_{D_1} = \omega - \omega_{D1}$, $\delta_{D_2} = \omega - \omega_{D2}$ is the frequency detuning of the laser. Therefore, according to equation (3) we only consider the scalar and vector shift in this work. We use tune-out wavelength for spin-dependent optical lattice, in which a.c. Stark shifts cancel. Two internal spin states have different tune-out wavelengths when the contributions of both the scalar and vector shifts are included[35].

We choose two ground hyperfine Zeeman states of ^{87}Rb $|F=2, m_F=0\rangle$ of the $F=2$, and $|F=1, m_F=1\rangle$ of the $F=1$ hyperfine manifold as the two internal spin states. A bias magnetic field with 10 Gauss is applied along the 45° diagonal line of the square lattice V_2. We scan the wavelength of the optical lattice beams to determine the tune-out wavelength precisely, as shown in Extended Data Fig. 2. The tune-out wavelength for $|1, 1\rangle$ state is determined at 788.28 nm with σ_- circular polarization as shown in Extended Data Fig. 2c, which balances the contribution of the scalar and vector shift. Thus, we choose this wavelength for the lattice V_2. Note that the tune-out wavelength for $|1, 1\rangle$ state is sensitive to the intersection angle between the laser beam wave vector and magnetic field quantization axis, which requires a careful alignment of the bias magnetic field. The spin state $|2, 0\rangle$ only experiences the square lattice V_2 with the red-detuning a.c. stark shift (which is only from scalar shift), as shown in Extended Data Fig. 2d,f; by contrast, the spin state $|1, 1\rangle$ experiences no shift.

On the other hand, there is only the contribution of the scalar shift for the spin state $|2, 0\rangle$; the tune-out wavelength for $|2, 0\rangle$ state is 790.02 nm as shown in Extended Data Fig. 2a, which is well known and studied experimentally[34,35]. We choose this tune-out wavelength of 790.02 nm with σ_+ circular polarization as the wavelength of the lattice V_1. Thus, the spin state $|1, 1\rangle$ experiences the square lattice V_1 with the red-detuned a.c. stark shift. By contrast, the spin state $|2, 0\rangle$ sees zero light shift. Note that the tune-out wavelength for $|2, 0\rangle$ state is insensitive to the intersection angle between the laser beam wave vector and magnetic field quantization axis. The spin state $|1, 1\rangle$, however, has a different lattice depth in two orthogonal directions of the lattice V_1, respectively, and feels the lattice V_1 with the red-detuning a.c. stark shift (which is only from vector shift at the wavelength of 790.02 nm) as shown in Extended Data Fig. 2b,e.

A moiré superlattice can be generated by a small difference in lattice constant or orientation. Because two different wavelengths are used for twisted-bilayer lattices in this work, there is a large-period superlattice with $\Delta\lambda = 179$ µm, much larger than the size of atomic cloud. Therefore, we can adjust the retroreflected concave mirror to load atoms into the lower potential well of the long-period superlattice and neglect the influence on the measurement of moiré pattern. In the future, we can correct this effect of two different wavelengths by using a slight angle lattice beam for V_2 to ensure the same lattice constant for two lattice potentials.

Band structures and flat band

For square lattices, the commensurate angles θ satisfy $\tan(\theta/2) = \overline{m}/\overline{n}$, where $\overline{m}$ and $\overline{n}$ are coprime natural number. An equivalent condition is $\cos\theta = \overline{a}/\overline{c}$ and $\sin\theta = \overline{b}/\overline{c}$, which can be defined by Pythagorean triples ($\overline{a}^2 + \overline{b}^2 = \overline{c}^2$, where $(\overline{a}, \overline{b}, \overline{c}) \in N$ are positive integers)[13]. The relationship between $(\overline{m}, \overline{n})$ and $(\overline{a}, \overline{b}, \overline{c})$ is $(\overline{m} + i\overline{n})^2 = (\overline{a} + i\overline{b})$ when $(\overline{m} + \overline{n}) \in$ odd and $(\overline{m} + i\overline{n})^2 = 2(\overline{a} + i\overline{b})$ when $(\overline{m} + \overline{n}) \in$ even. For the commensurate optical lattice, the period of its supercell (sc) is given by $\lambda_{sc} = \overline{m}a/\sin(\theta/2) = 2\overline{m}\lambda_{mo}$ when $(\overline{m} + \overline{n}) \in$ odd and $\lambda_{sc} = \overline{m}a/\sin(\theta/2) = \sqrt{2}\,\overline{m}\lambda_{mo}$ when $(\overline{m} + \overline{n}) \in$ even.

Here, we choose the band structure of the commensurate optical lattice with the commensurate angle $\theta = 2\arctan(1/22)$ as an approximation of the experimental case. If getting a better approximation of band structure for the experimental case, we can choose the larger supercell to calculate the energy band structure, whose commensurate angle is closer to the experimental case. The band structure $E(\mathbf{k})$ of the commensurate optical lattice can be obtained by solving the stationary Schrödinger equation, $H\Psi = E\Psi$, with the Bloch function, $\Psi(\mathbf{r}) = \exp^{i\mathbf{k}\cdot\mathbf{r}} u(\mathbf{r})$. Here the Hamiltonian H is given as

$$H = \begin{pmatrix} -\frac{\hbar^2}{2m}\nabla^2 + V_1 + \frac{\Delta}{2} & \Omega_R \\ \Omega_R & -\frac{\hbar^2}{2m}\nabla^2 + V_2 - \frac{\Delta}{2} \end{pmatrix}, \quad (4)$$

$u(\mathbf{r})$ is a periodic function with the same periodicity as the coupled lattice. The spin-dependent square optical lattice with a twist angle θ can be described by the potentials

$$\begin{aligned} V_1 &= V_0\left[\sin^2\left(kx\cos\frac{\theta}{2} - ky\sin\frac{\theta}{2}\right) + \sin^2\left(ky\cos\frac{\theta}{2} + kx\sin\frac{\theta}{2}\right)\right], \\ V_2 &= V_0\left[\sin^2\left(kx\cos\frac{\theta}{2} + ky\sin\frac{\theta}{2}\right) + \sin^2\left(ky\cos\frac{\theta}{2} - kx\sin\frac{\theta}{2}\right)\right], \end{aligned} \quad (5)$$

where $k = 2\pi/\lambda$ is the wave number of lasers for the lattice and V_0 describes the lattice depth. In numerics, we first discretize the unit supercell of area $\sqrt{\overline{c}}a \times \sqrt{\overline{c}}a$ in real space ($\overline{c}$ is the largest value in the Pythagorean triple) into $l \times l$ grids, and then diagonalize the effective Hamiltonian for $u(\mathbf{r})$. As shown in Extended Data Fig. 3, the band structure approaches the flat band when increasing the interlayer coupling.

As our system allows for flexible control of the interlayer couplings, the flat band in the lowest energy band can be realized. The Hamiltonian equation (4) can be formally diagonalized as

$$H = \begin{pmatrix} H^+_{\mathrm{eff}} & 0 \\ 0 & H^-_{\mathrm{eff}} \end{pmatrix}, \quad (6)$$

where

$$H^{\pm}_{\mathrm{eff}} = h_0 \pm h_1, \quad (7)$$

with $h_0 = -\frac{\hbar^2}{2m}\nabla^2 + \frac{V_1 + V_2}{2}$, and $h_1 = \sqrt{\Omega_R^2 + \frac{(V_1 - V_2 + \Delta)^2}{4}}$. In the large interlayer coupling limit, $\Omega_R \gg V_0, \Delta$, the low-energy band structure is encoded in the effective Hamiltonian H^-_{eff} in the lower-right block, which can be further approximated as

$$H^-_{\mathrm{eff}} \approx -\frac{\hbar^2}{2m}\nabla^2 + \frac{V_1 + V_2}{2} - \frac{(V_1 - V_2 + \Delta)^2}{8\Omega_R} - \Omega_R, \quad (8)$$

or in a rougher way

$$H^-_{\mathrm{eff}} \approx -\frac{\hbar^2}{2m}\nabla^2 + \frac{V_1 + V_2}{2} - \Omega_R, \quad (9)$$

The approximated effective Hamiltonians correspond to some effective lattices for a single-layer (single-component) system, separately, $V = \frac{V_1 + V_2}{2} - \frac{(V_1 - V_2 + \Delta)^2}{8\Omega_R}$ for equation (8), and $V = \frac{V_1 + V_2}{2}$ for equation (9), with certain global energy shift. Specifically, equation (9) indicates that the system becomes the single layer (single-component) experiencing a twisted optical lattice.

When increasing the interlayer coupling into the strong region, the long-wavelength moiré potential becomes deeper, so atoms in the lowest band are isolated on a larger spatial scale (moiré wavelength), which enhances the wave function localization and contributes to the creation of a flat band. The single-layer system with a twisted optical lattice (approximation at the strong interlayer coupling limit) admits a flat-band structure in the lowest band, which has been studied experimentally in photonic system[13–15]. The moiré flat bands have several advantages.

First, the flat bands quench kinetic energy scales (wave function localization), thereby drastically enhance the role of interactions and amplify the effects of interactions. Second, the moiré superlattice leads to the emergence of minibands within a reduced Brillouin zone. The small Brillouin zone means that low atomic densities are sufficient for full filling or depletion of the superlattice bands, which is easily controlled in an experiment.

Theoretical calculation of the modified superfluid to insulator transition

In the mean-field approximation, the system for the superfluid phase can be well described by the coupled Gross–Pitaevskii (GP) equations

$$\begin{aligned} i\hbar\frac{\partial\psi_1}{\partial t} &= \left[-\frac{\hbar^2}{2m}\nabla^2+\frac{1}{2}m\omega_\perp^2(x^2+y^2)+V_1+\eta g_{11}\,|\psi_1|^2+\eta g_{12}\,|\psi_2|^2\right]\\ &\quad \psi_1+\hbar\Omega_\mathrm{R}\psi_2,\\ i\hbar\frac{\partial\psi_2}{\partial t} &= \left[-\frac{\hbar^2}{2m}\nabla^2+\frac{1}{2}m\omega_\perp^2(x^2+y^2)+V_2+\eta g_{12}\,|\psi_1|^2+\eta g_{22}\,|\psi_2|^2\right]\\ &\quad \psi_2+\hbar\Omega_\mathrm{R}\psi_1, \end{aligned} \tag{10}$$

where the MW detuning is $\Delta = 0$ and the wave function is normalized as $\sum_i\int|\psi_i|^2 d\mathbf{r} = N$, with N the total atom number. The strong confinement along the z axis gives rise to the quasi-2D interaction strengths represented by a reduction coefficient η multiplied by $g_{ij} = 4\pi\hbar^2 a_{ij}/m$, where $\eta^{-1} = \sqrt{h/m\omega_z}$ defines the characteristic length along the z axis and a_{ij} is the 3D s-wave scattering length. In the experiment, the trapping frequency $\omega_z \cong 2\pi\times 1$ kHz, and the scattering length for the ^{87}Rb atoms is about $a_{ij} \cong 100a_\mathrm{B}$ with a_B the Bohr radius. This indicates that even though the system is thermodynamically 2D, the collisions still keep their 3D character with $\eta^{-1} \gg a_{ij}$. Considering the similarity in scattering lengths a_{11}, a_{22} and a_{12} for the ^{87}Rb atoms, in the calculation we focus on the SU(2) symmetric interaction with $g = g_{11} = g_{22} = g_{12}$. In addition to the intercomponent atomic interaction, the two components are also coupled by a MW pulse, which causes Rabi oscillations with the frequency Ω_R.

By using the imaginary time evolution method, one can solve the GP equations numerically for the ground states in the harmonic trap. Theoretically, the non-commensurate twist angle $\theta = 5.21°$ should be a localized single particle ground state whereas the commensurate angle $\theta = 2\arctan\frac{1}{22}$ gives rise to extended ground states in the absence of interactions. Experimentally, the interatomic interaction is dominant, and always leads to extended many-body states with the aperiodic and periodic bilayer lattices becoming almost indistinguishable.

The phase transition from superfluid to MI can be well described by the Bose–Hubbard model in the tight-binding approximation. For simplicity we consider the interlayer coupling as a quasi-periodically perturbed potential, which leads to a site-dependent energy offset

$$M_i = M_\mathrm{R}\,[\sin^2(i_x\pi\cos\theta + i_y\pi\sin\theta) + \sin^2(i_y\pi\cos\theta - i_x\pi\sin\theta)], \tag{11}$$

where the subindex i_x and i_y label the position of the ith site in the two-dimensional space. The tight-binding Hamiltonian for one layer then is given by

$$H = -t\sum_{\langle i,j\rangle} b_i^\dagger b_j + \frac{U}{2}\sum_i \hat{n}_i(\hat{n}_i-1) + \sum_i (M_i-\mu)\hat{n}_i, \tag{12}$$

where the first term describes the nearest-neighbour tunnelling with $b^\dagger$ and b being the creation and annihilation operators, and the second term represents the on-site interaction. The hopping amplitude t is considered to be site-independent for weak interlayer coupling and can be estimated by $t = \frac{4}{\sqrt{\pi}}E_\mathrm{r}(V_0/E_\mathrm{r})^{3/4}\mathrm{e}^{-2(V_0/E_\mathrm{r})^{1/2}}$. The local repulsion U depends on the depth of the optical lattice, and is given by $U = \sqrt{8/\pi}\,ka_sE_\mathrm{r}(V_0/E_\mathrm{r})^{3/4}$ (ref. [46]). The chemical potential μ controls the average number of atoms in the moiré lattice.

The mean-field phase diagram (Fig. 5a) can be mapped by using the Gutzwiller method, which expands the local state $|\psi_i\rangle$ at site i in the Fock basis[47,48]. When the interlayer coupling $M_i = 0$, the system is reduced to the standard Bose–Hubbard model[38], which includes two phases, the superfluid phase and the MI phase[44,49]. While the superfluid phase is identified by the superfluid order parameter $\langle\hat{b}_i\rangle \neq 0$ and an arbitrary filling of the atoms on the site, the MI phase emerges with an integer number of atoms per site with $\langle\hat{b}_i\rangle = 0$. When the interlayer coupling $M_i \neq 0$, the persistent coherence of the moiré and primary lattice length scale as well as density distribution in real space can be used to distinguish the phases, which is determined by the order parameter $\langle\hat{b}_i\rangle$ and the filling of the atoms on the site n as shown in Extended Data Fig. 4. The chemical potential $\mu/U = 1$ is considered in this calculation.

Phase diagram with zero temperature and the homogeneous system is predicted theoretically as shown in Extended Data Fig. 4a, in which four phases of superfluid (SF), superfluid II (SF-II), MI and insulator (I) are included. The SF-II phase is a state with superfluid domains embedded in a gapped insulate state, which is caused by interlayer coupling. So, the SF-II phase can be identified by checking the disappearance of the moiré-scale long-range correlation with vanishing moiré lattice momentum but the short-range coherence with residual primary lattice momentum remains. At the same time, the SF-II phase supports the moiré pattern in the real space. As MI is an incompressible insulator for integer filling factor with a gap U for particle-hole excitations induced by the on-site interaction U, the moiré pattern in the real space appears only when the interlayer coupling strength is larger than U to break this gap. Therefore, the I phase supports the moiré pattern in the real space and no spatial coherence at all scales, which approaches the MI phase without the moiré pattern in the real space in the limit of weak interlayer coupling. Here, the phase transition between the SF to the MI phase should have an intermediate phase SF-II. Obviously, the coherence is lost almost simultaneously in all length scales at the critical point at very weak interlayer coupling only for zero temperature as shown in Fig. 5a. By contrast, at finite temperature, the thermodynamic quantities behave smoothly near the critical point and more long-range moiré coherence is lost before the short-range primary lattice length. Therefore, there exists an SF-MI critical regime, which seems likely to be a thermodynamic phase and is not predicted theoretically at zero temperature. The SF-MI critical regime has the same coherence as SF-II without the moiré-scale long-range correlation but with the short-range coherence. However, the SF-MI critical regime does not support the moiré pattern in real space that is distinguished from SF-II. In fact, the insulator phase presents special characteristics due to the strong interaction and quasi-disorder induced by the larger interlayer coupling, similar to a Bose glass insulator from the model of a disordered strongly interacting bosonic system[38–40].

Data availability

All data generated or analysed during this study are included in this published article. Further data are also available from the corresponding authors upon reasonable request.

43. Xiong, D., Wang, P., Fu, Z., Chai, S. & Zhang, J. Evaporative cooling of ^{87}Rb atoms into Bose-Einstein condensate in an optical dipole trap. *Chin. Opt. Lett.* **8**, 627–629 (2010).
44. Greiner, M., Mandel, O., Esslinger, T., Hänsch, T. W. & Bloch, I. Quantum phase transition from a superfluid to a Mott insulator in a gas of ultracold atoms. *Nature* **415**, 39 (2002).
45. Steck, D. A. *Quantum and Atom Optics* https://atomoptics.uoregon.edu/~dsteck/teaching/quantum-optics/ (2007).
46. Zwerger, W. Mott Hubbard transition of cold atoms in optical lattices. *J. Opt. B: Quantum Semiclass. Opt.* **5**, S9–S16 (2003).
47. Krauth, W., Caffarel, M. & Bouchaud, J. P. Gutzwiller wave function for a model of strongly interacting bosons. *Phys. Rev. B* **45**, 3137–3140 (1992).

48. Sheshadri, K., Krishnamurthy, H. R., Pandit, R. & Ramakrishnan, T. V. Superfluid and insulating phases in an interacting-Boson model: mean-field theory and the RPA. *Europhys. Lett.* **22**, 257–263 (1993).
49. Jaksch, D., Bruder, C., Cirac, J. I., Gardiner, C. W. & Zoller, P. Cold Bosonic atoms in optical lattices. *Phys. Rev. Lett.* **81**, 3108–3111 (1998).

Acknowledgements This research is supported by Innovation Program for Quantum Science and Technology (grant no. 2021ZD0302003), National Key Research and Development Program of China (grant nos. 2016YFA0301602, 2018YFA0307601 and 2022YFA1404101), NSFC (grant nos. 12034011, 12022406, 12074342 and 11804203), the Fund for Shanxi '1331 Project' Key Subjects Construction and Tencent (Xplorer Prize). C.C. acknowledges support by the National Science Foundation (grant no. PHY-2103542) and the Army Research Office STIR (grant no. W911NF2110108).

Author contributions J.Z. conceived the idea and performed the experimental designs. L.W., Z.M., F.L., K.W., P.W. and J.Z. performed the experiments. C.C., Z.M., L.W., F.L., W.H. and J.Z. analysed the data and all authors discussed the results. W.H., C.G. and J.Z. performed the simulation. Z.M. plotted the figures. J.Z. and C.C. wrote the manuscript. All authors interpreted the results and reviewed the manuscript. J.Z. designed and supervised the project.

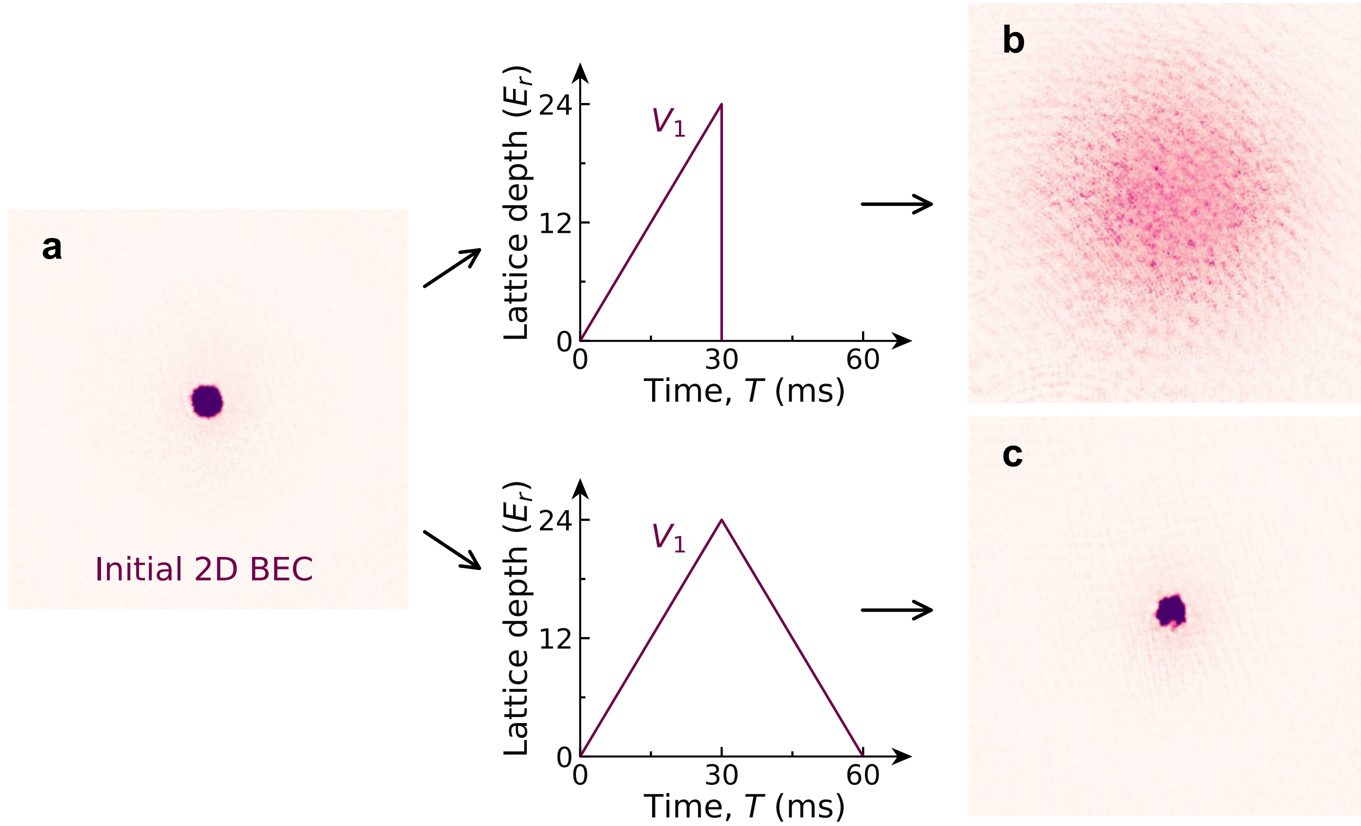
Competing interests The authors declare no competing interests.

Additional information

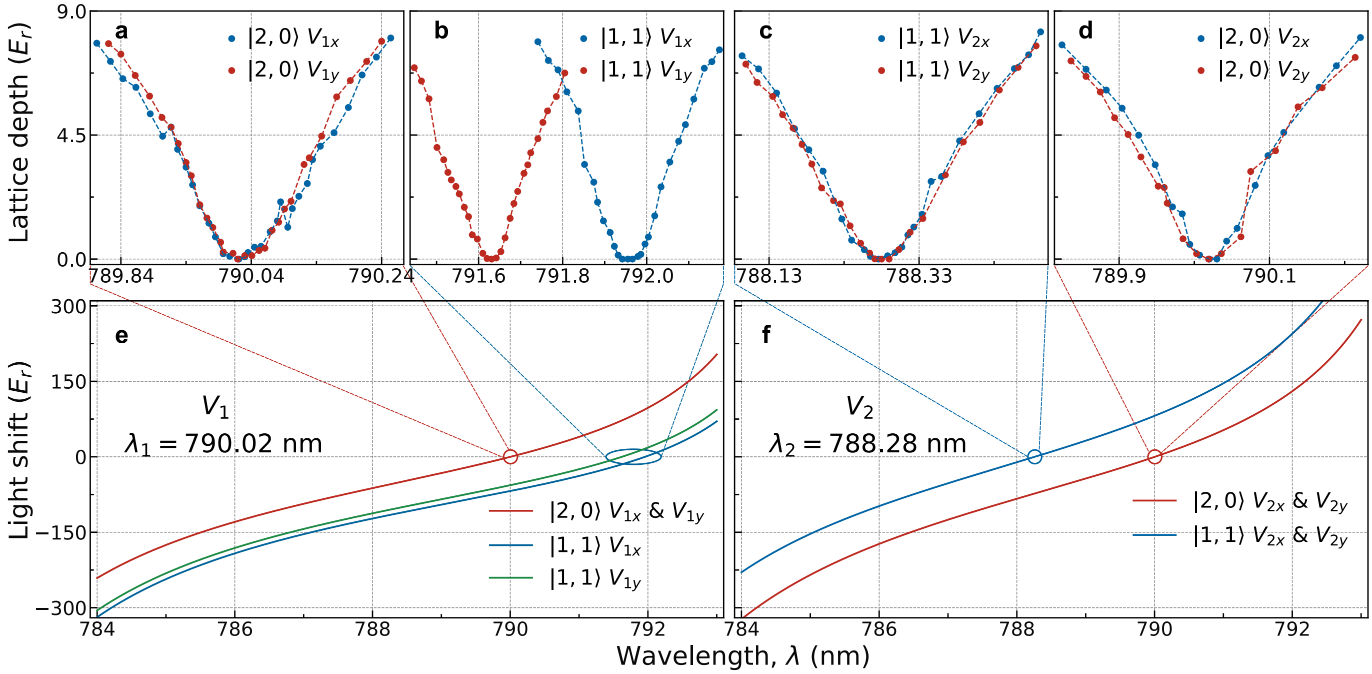
Correspondence and requests for materials should be addressed to Jing Zhang.

Peer review information *Nature* thanks Richard Schmidt and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

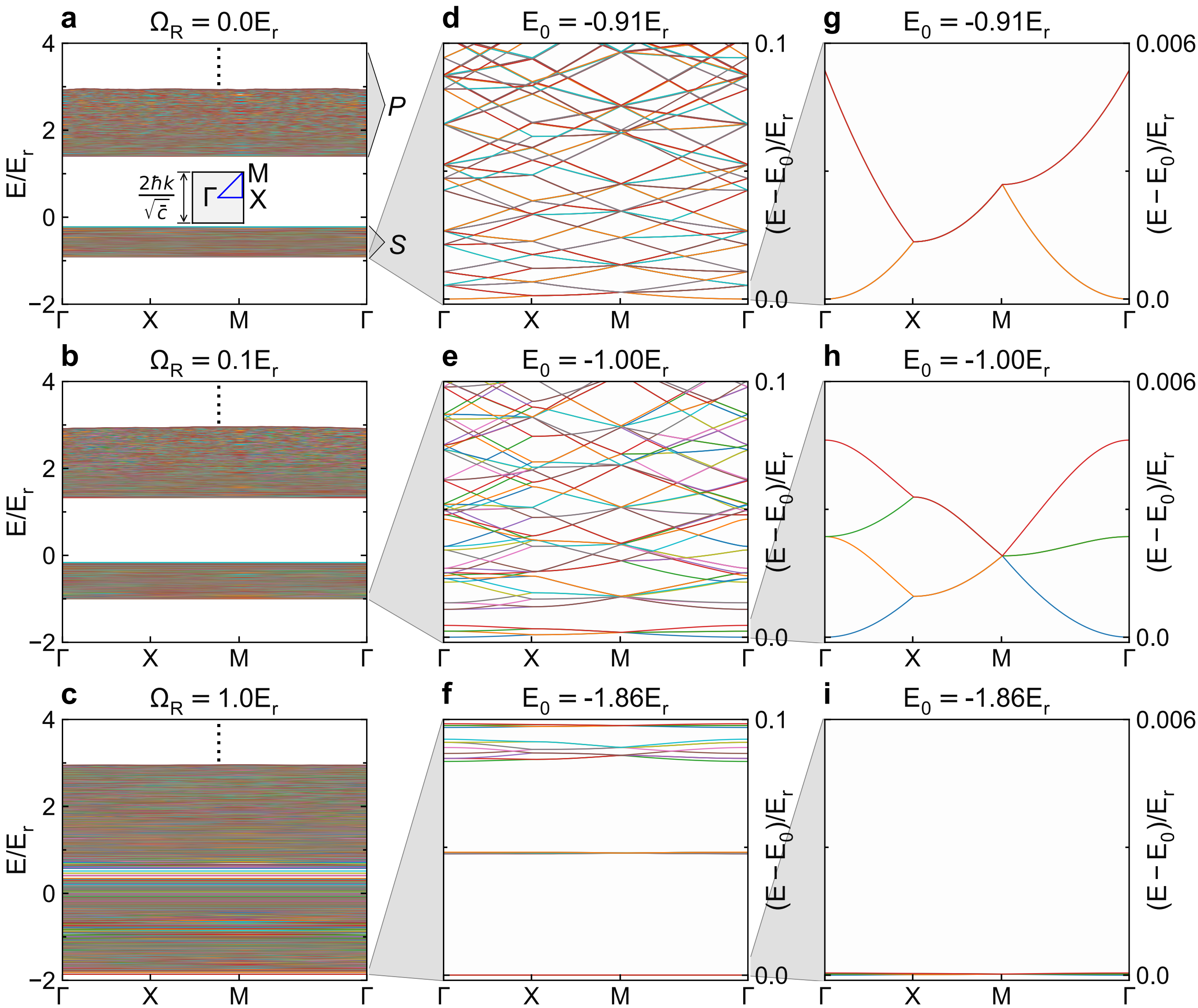
Reprints and permissions information is available at http://www.nature.com/reprints.



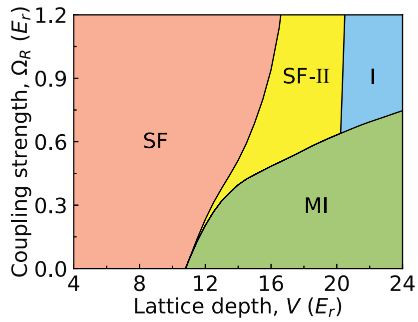
Extended Data Fig. 1 | Coherence in the SF-MI transition. **a**, The initial BEC in 2D pancake-like potential. **b**, Absorption image after atoms are released abruptly from an optical lattice potential V_1 (or V_2) with a potential depth $24E_r$. **c**, Absorption image when the lattice is ramped up to the lattice depth $24E_r$ and then ramp down to zero. The images are obtained after 18 ms free space expansion.



Extended Data Fig. 2 | Determination of tune-out wavelengths. a–b, The lattice depth V_{1x} (blue) and V_{1y} (red) as a function of wavelength λ for the two different hyperfine states $|F=1, m_F=1\rangle$ and $|F=2, m_F=0\rangle$. The angles between V_{1x}, V_{1y} and B_0 are 39.79° and 50.21° respectively. **c–d**, The potential depth V_{2x} (blue) and V_{2y} (red) as a function of wavelength λ for the two different hyperfine states $|F=1, m_F=1\rangle$ and $|F=2, m_F=0\rangle$. **e**, Theoretical light shift of V_{1x}, V_{1y} for $|1, 1\rangle$ and $|2, 0\rangle$. **f**, Theoretical lattice depth of V_{2x}, V_{2y} for $|1, 1\rangle$ and $|2, 0\rangle$. The bias magnetic field of 10 Gauss is applied along the 45° diagonal line of the square lattice V_2.

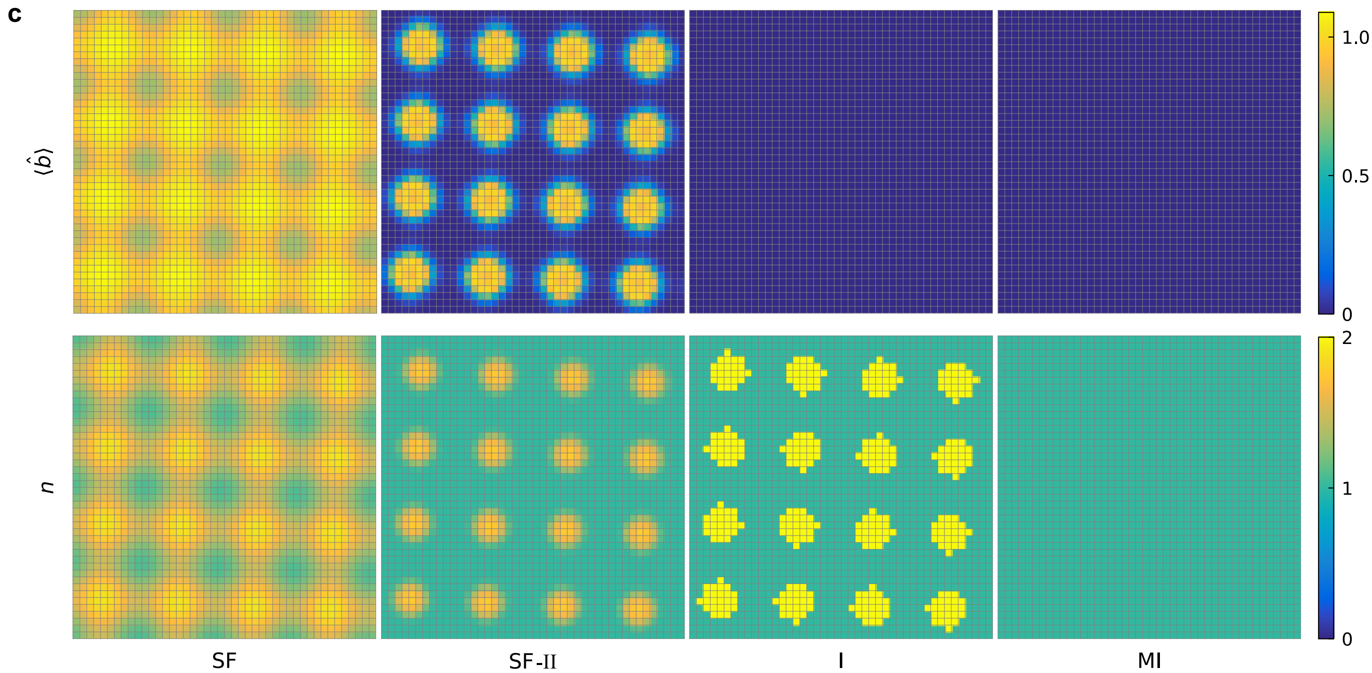


Extended Data Fig. 3 | Band structure of the twisted-bilayer optical lattices. The twist angle of the commensurate optical lattice is $\theta = 2\arctan(1/22)$, whose band structure is regarded as an approximation of the experimental case $\theta = 5.21°$. **a**, **b** and **c** show the band structures for the interlayer coupling strength $\Omega_R = 0E_r$, $0.1E_r$ and $1E_r$ respectively. **a** also gives the band structure without the interlayer coupling in the form of the superlattice minibands within the same reduced Brillouin zone. **d**,**e** and **f** are the enlargement of the lowest bands of **a**,**b**, and **c**, respectively. **g**,**h** and **i** are the further enlargement of the lowest bands of **d**,**e** and **f**, respectively. Here, the MW detuning is $\Delta = 0$, $V_0 = 4E_r$ and E_0 corresponds to the energy of the lowest band.



b

	Real space	Momentum space	
	Moiré pattern	Coherence at the primary lattice length-scale	Coherence at the moiré length-scale
SF	Yes	Yes	Yes
SF-II	Yes	Yes	No
I	Yes	No	No
MI	No	No	No



Extended Data Fig. 4 | Characteristics of the different phases. a, Phase diagram, where SF, SF-II, MI, and I refer to superfluid, superfluid only with short-range coherence, Mott insulator, and insulator. **b**, Table shows the features of the different phases. **c**, Plots of the order parameter $\langle \hat{b}_i \rangle$ and the filling of the atoms on the site n for the different phases. Parameters (V/E_r, Ω_R/E_r) are (10,0.6), (15,0.6), (23,0.3) and (23,1.1) for the plots from left to right respectively. The chemical potential $\mu/U = 1$ is considered.

Article

Geometric frustration of Jahn–Teller order in the infinite-layer lattice

https://doi.org/10.1038/s41586-022-05681-2

Received: 5 April 2022

Accepted: 22 December 2022

Published online: 22 February 2023

Check for updates

Woo Jin Kim[1,2]✉, Michelle A. Smeaton[3], Chunjing Jia[1,4], Berit H. Goodge[5,6], Byeong-Gwan Cho[7], Kyuho Lee[1,8], Motoki Osada[1,9], Daniel Jost[1], Anton V. Ievlev[10], Brian Moritz[1], Lena F. Kourkoutis[5,6], Thomas P. Devereaux[1,9] & Harold Y. Hwang[1,2]✉

The Jahn–Teller effect, in which electronic configurations with energetically degenerate orbitals induce lattice distortions to lift this degeneracy, has a key role in many symmetry-lowering crystal deformations[1]. Lattices of Jahn–Teller ions can induce a cooperative distortion, as exemplified by $LaMnO_3$ (refs. [2,3]). Although many examples occur in octahedrally[4] or tetrahedrally[5] coordinated transition metal oxides due to their high orbital degeneracy, this effect has yet to be manifested for square-planar anion coordination, as found in infinite-layer copper[6,7], nickel[8,9], iron[10,11] and manganese oxides[12]. Here we synthesize single-crystal $CaCoO_2$ thin films by topotactic reduction of the brownmillerite $CaCoO_{2.5}$ phase. We observe a markedly distorted infinite-layer structure, with ångström-scale displacements of the cations from their high-symmetry positions. This can be understood to originate from the Jahn–Teller degeneracy of the d_{xz} and d_{yz} orbitals in the d^7 electronic configuration along with substantial ligand–transition metal mixing. A complex pattern of distortions arises in a $2\sqrt{2} \times 2\sqrt{2} \times 1$ tetragonal supercell, reflecting the competition between an ordered Jahn–Teller effect on the CoO_2 sublattice and the geometric frustration of the associated displacements of the Ca sublattice, which are strongly coupled in the absence of apical oxygen. As a result of this competition, the $CaCoO_2$ structure forms an extended two-in–two-out type of Co distortion following 'ice rules'[13].

The cooperative Jahn–Teller effect (JTE) often induces strong coupling between charge, orbital and magnetic ordering, which has been related to many correlated phenomena such as colossal magnetoresistance[14,15] and superconductivity[16]. For the magnetoresistance associated with the metal–insulator transition in perovskite-based manganites, it has been established that the double-exchange model for magnetism cannot solely account for the large resistivity drop below the paramagnetic-to-ferromagnetic phase transition, and rather that a strong electron–phonon interaction arising from the JTE plays a crucial role[17–20]. Moreover, single-layered perovskite $(La,A)_2CuO_4$ (where A is Sr or Ba) shows unconventional superconductivity within Jahn–Teller (JT)-distorted CuO_6 octahedra, indicating the important role of the JTE in many correlated ground states[21].

Despite its generality, the study of the cooperative JTE in oxides has been largely confined to octahedral and tetrahedral coordination. In perovskites, octahedral-site cations having outer-electron configurations of d^4, d^7 or d^9 (ref. [4]) are often JT-active systems, splitting the partially filled e_g orbitals (Fig. 1a). In $LaMnO_3$ with high-spin state d^4, the partially filled e_g orbitals of Mn^{3+} induces both MnO_6 octahedral distortion (that is, the Q_3 JT distortion)[1] and d_{z^2}-orbital ordering to reduce its total energy (Fig. 1b), resulting in an orthorhombic crystal structure (Fig. 1c). The cooperative JTE has also been observed in spinel structures with tetrahedral-site cations having d^1 or d^9 configurations[5]. The JTE in a square-planar anion configuration has been theoretically discussed from the foundations of the concept: that a Q_2 vibrational mode will distort the high-symmetry square coordination to a low-symmetry rhombus[1]. The FeO_4 chromophore molecule exhibits nearly square-planar coordination, which is thought to be JT driven from the expected tetrahedral structure[22]. However, a pure cooperative Q_2 JT distortion has not been experimentally observed in crystalline materials hosting square-planar-coordinated cations.

The infinite-layer structure provides a unique platform to explore this possibility and the potential cooperative response in a lattice of JT ions. This structure forms a four-coordinated transition metal oxide with a two-dimensional (2D) square lattice, which can often be derived from precursor higher oxides by de-intercalating the apical oxygen using topotactic reduction (Fig. 2a). These 2D oxides were first synthesized in powder nickelates with an extremely low nickel valence state (Ni^{+1}) (ref. [8]). The same four-coordinated system has also been found in copper oxide superconductors[6,7]. Subsequently, the iron oxide infinite-layer structure was discovered with antiferromagnetic ground states[10,11]. In addition, recently, hole-doped infinite-layer nickelates with Ni $3d^{9-\delta}$ have been found to be superconducting[23,24].

Within this quasi-2D structure, how the orbital degeneracy affects the crystal and electronic structure is an open experimental question.

[1]Stanford Institute for Materials and Energy Sciences, SLAC National Accelerator Laboratory, Menlo Park, CA, USA. [2]Department of Applied Physics, Stanford University, Stanford, CA, USA. [3]Department of Materials Science and Engineering, Cornell University, Ithaca, NY, USA. [4]Department of Physics, University of Florida, Gainesville, FL, USA. [5]School of Applied and Engineering Physics, Cornell University, Ithaca, NY, USA. [6]Kavli Institute at Cornell for Nanoscale Science, Cornell University, Ithaca, NY, USA. [7]Pohang Accelerator Laboratory, POSTECH, Pohang, Republic of Korea. [8]Department of Physics, Stanford University, Stanford, CA, USA. [9]Department of Materials Science and Engineering, Stanford University, Stanford, CA, USA. [10]Center for Nanophase Materials Sciences Oak Ridge National Laboratory, Oak Ridge, TN, USA. ✉e-mail: wjk316@stanford.edu; hyhwang@stanford.edu

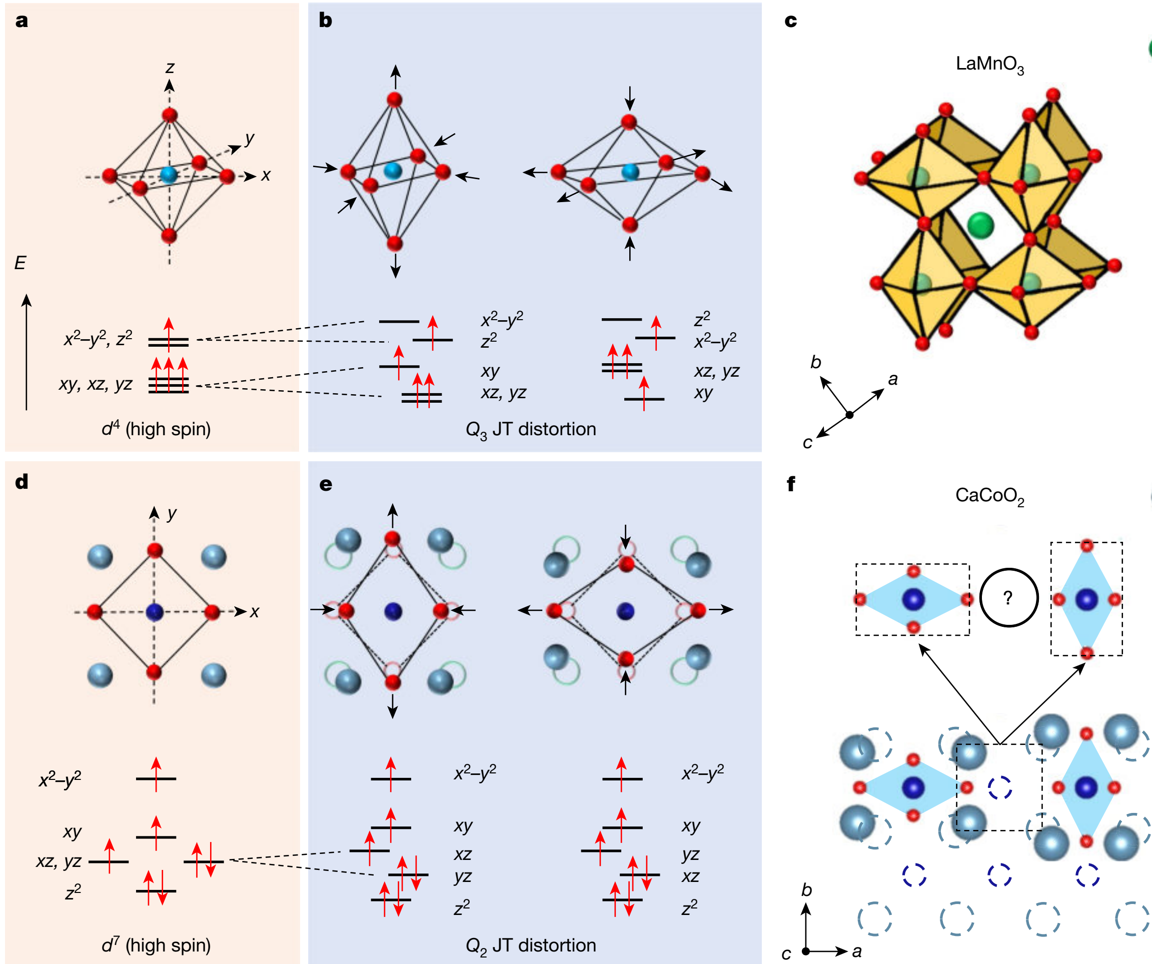


Fig. 1 | JT distortion in 3D and 2D oxide lattices. a, Schematic MO_6 (where M is a transition metal) octahedron and crystal field structure for a high-spin d^4 electron configuration. **b**, Octahedral distortions and associated orbital levels for a Q_3 JT distortion. **c**, Strong Q_3 JT-distorted $LaMnO_3$ crystal structure with high-spin d^4 Mn^{3+} as an example of a cooperative JT distortion. **d**, Schematic square-planar MO_4 plaquette and crystal field structure for a high-spin d^7 electron configuration. **e**, Square-planar distortion and associated orbital levels for a Q_2 JT distortion. **f**, Displacements in the Ca layer in $CaCoO_2$, arising from strong interlayer coupling in the absence of apical oxygen, geometrically frustrate a simple cooperative JT distortion analogous to $LaMnO_3$.

The crystal field structure for the square-planar coordination is quite different from that of octahedral or tetrahedral coordination, with lower overall degeneracy. As shown in Fig. 1d, only the d_{xz} and d_{yz} orbitals are degenerate. Therefore, to be JT active, the system should have a partially filled d_{xz} (or d_{yz}) orbital state. The high-spin d^7 configuration meets these conditions, making the infinite-layer cobaltates (stabilized here in $CaCoO_2$) with Co^{2+} oxidation state an ideal candidate. Considering only the CoO_2 plane, one might expect a cooperative JTE to induce an orbitally ordered lattice of d_{xz} or d_{yz} orbital states in analogy to $LaMnO_3$ (Fig. 1e). However, the absence of apical oxygen gives rise to strong electrostatic coupling between the CoO_2 and Ca layers—an interaction that is largely screened for octahedral coordination. As we will show, this interlayer coupling gives rise to a geometric frustration of the induced Ca displacements (Fig. 1f), which is in competition with the cooperative JTE.

Ångström-scale cation displacements in $CaCoO_2$

Here we stabilized a infinite-layer compound $CaCoO_2$ with dramatic in-plane lattice distortions. The starting point is the synthesis of epitaxial single-crystal thin films of brownmillerite $CaCoO_{2.5}$ on strontium titanium oxide ($SrTiO_3$) (001) substrates (Methods; θ– diffraction angle (2θ) symmetric X-ray diffraction (XRD) scan in Fig. 2b)[25]. The high-angle annular dark-field (HAADF) scanning transmission electron microscopy (STEM) image of $CaCoO_{2.5}$ (Extended Data Fig. 1a) confirms the excellent crystallinity with abrupt interfaces with the substrate and subsequent $SrTiO_3$ capping layer, which is used to stabilize the reduced phase. The $CaCoO_{2.5}$ heterostructure was then reduced using calcium hydride (CaH_2), after which the XRD symmetric scan shows only peaks corresponding to the infinite-layer structure (00*l*) (Fig. 2c), with a substantial contraction of the out-of-plane lattice parameter (c_t) from about 3.72 Å to about 3.27 Å. From reciprocal space mapping, we extract the in-plane lattice constant (a_t) of $CaCoO_2$ to be about 3.86 Å (Extended Data Fig. 1b), indicating that the reduced film is relaxed from the substrate.

The large *c*-axis difference between $CaCoO_2$ and $CaCoO_3$ (ref. [26]) closely follows trends of known infinite-layer oxides (Extended Data Fig. 1c). Furthermore, electron energy loss spectroscopy (EELS) of the $CaCoO_{2.5}$ and $CaCoO_2$ films (Extended Data Fig. 2) is consistent with a transition from Co^{3+} to Co^{2+} (ref. [27]). It is noted that $SrCoO_2$ has been previously synthesized by reduction from $SrCoO_{2.5}$ (ref. [28]). However, $SrCoO_2$ has a much larger c_t (about 3.76 Å) with tetrahedrally coordinated Co^{2+}—that is, a fundamentally different structure[28]. Moreover, chemical characterization by time-of-flight secondary-ion mass spectrometry (TOF-SIMS) shows no sign of hydrogen intercalation in $CaCoO_2$ (Extended Data Fig. 3a), unlike structures such as $SrCoO_xH_y$ (refs. [29,30]). To our knowledge, the infinite-layer structure we observe has not previously been reported for cobalt oxides.

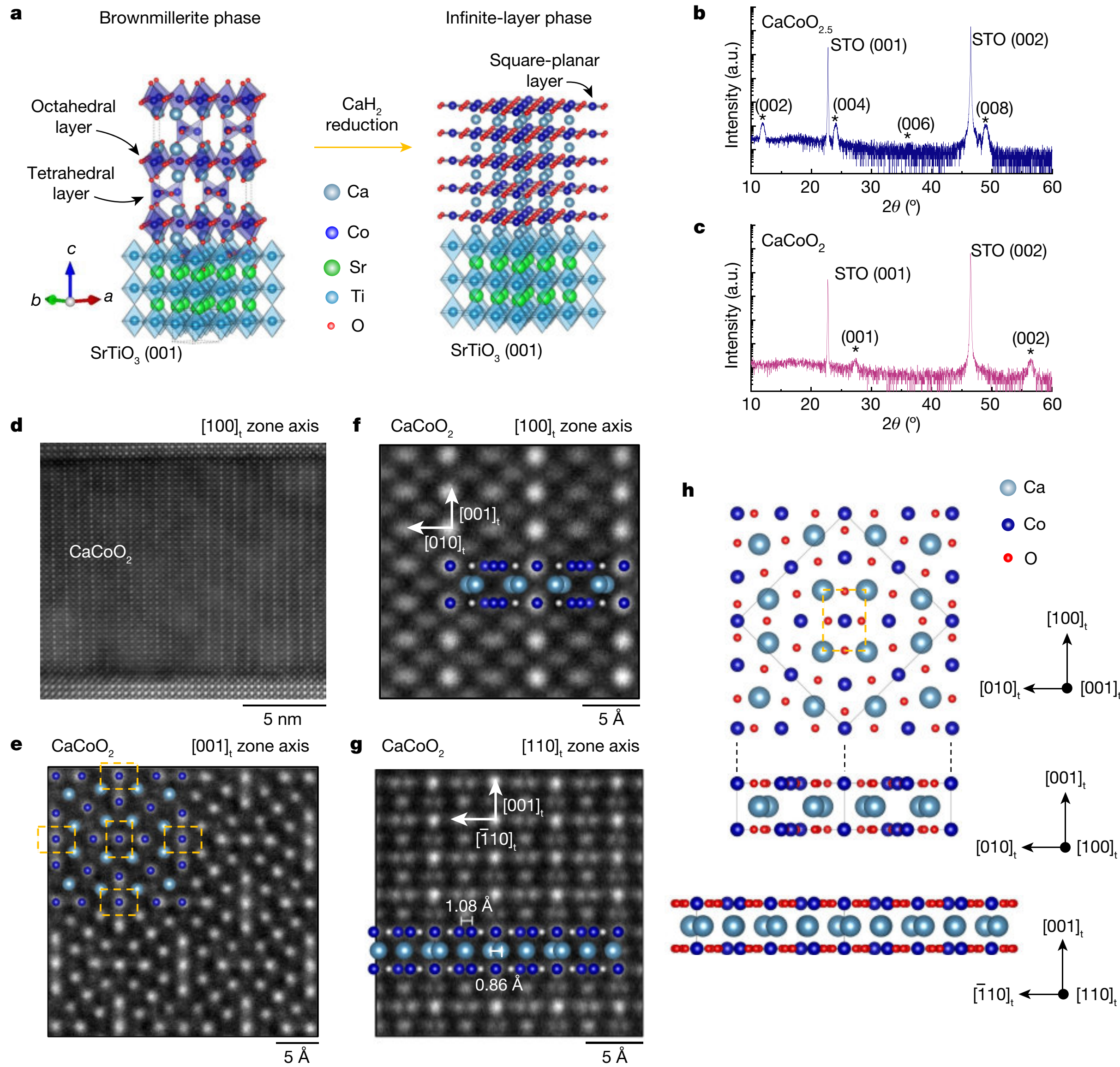


Fig. 2 | Synthesis and large-scale cation displacements in thin-film $CaCoO_2$. **a**, Schematic of the topotactic reduction of brownmillerite $CaCoO_{2.5}$ (left) to infinite-layer $CaCoO_2$ (right). **b**,**c**, XRD θ–2θ symmetric scans of 18-nm-thick $CaCoO_{2.5}$ (**b**) and $CaCoO_2$ (**c**) films capped with 2-nm-thick $SrTiO_3$ (STO) layers grown on $SrTiO_3$ (001) substrates. Asterisks indicate the diffraction peaks of the thin films. **d**, HAADF-STEM image of the $[100]_t$ zone axis of $CaCoO_2$. **e**, Plan-view HAADF-STEM image of $CaCoO_2$. The distorted Ca layer is clearly seen in the image (yellow dashed rectangles). **f**,**g**, High-magnification HAADF-STEM images of $CaCoO_2$ along the $[100]_t$ (**f**) and $[110]_t$ (**g**) zone axes. Dark (bright) blue balls indicate Co (Ca) cations and grey balls indicate oxygen. The atomic positions of oxygen here are shown for the ideal square-planar sites. **h**, Cation unit-cell structure of $CaCoO_2$ based on HAADF-STEM. The red solid balls indicate the oxygen positions after GIXRD refinement.

The HAADF-STEM image along the $SrTiO_3$ [100] zone axis (Fig. 2d) shows that $CaCoO_2$ has a very strong in-plane superlattice modulation, unlike all known infinite-layer systems[6,9,10,31]. A closer look at the $CaCoO_2$ layer along the $[100]_t$ zone axis in Fig. 2f shows that this is due to in-plane cation displacements from the simple tetragonal infinite-layer structure. These are more clearly visible along the $[110]_t$ zone axis (Fig. 2g), which reveals that both Co and Ca ions have alternating sites with very large splitting of 1.08 Å and 0.86 Å, respectively. Beyond these substantial in-plane distortions, there are no observable accompanying out-of-plane atomic distortions (Extended Data Fig. 3b–d). The θ–2θ asymmetric XRD scan also rules out a possible c-lattice doubling (Extended Data Fig. 4). The plan-view HAADF-STEM image in Fig. 3e also shows strong in-plane superlattice modulation. In aggregate, these results show that the cation lattice is described by a tetragonal supercell $2\sqrt{2}a_t \times 2\sqrt{2}a_t \times c_t$ ($a = b = 2\sqrt{2}a_t = 10.78$ Å and $c = c_t = 3.27$ Å) in the $P42_12$ symmetry group as shown in Fig. 2h (see Extended Data Table 1 for atomic coordinates). The projections of this crystal structure are shown in Fig. 2e–g, in close correspondence with the STEM images.

Structural refinement of the oxygen positions in $CaCoO_2$

Given the extremely small sample volume and low relative X-ray and electron scattering cross-sections for oxygen, determining their positions is quite challenging. Nevertheless, we can refine the oxygen positions from high-resolution synchrotron grazing incidence XRD (GIXRD) of the in-plane superlattice diffraction peaks (Fig. 3a). Using the cation coordinates from STEM, we first simulated the XRD pattern to find the seven strongest in-plane superlattice peaks (Extended Data Table 2) assuming undistorted oxygen positions. These were then used to experimentally

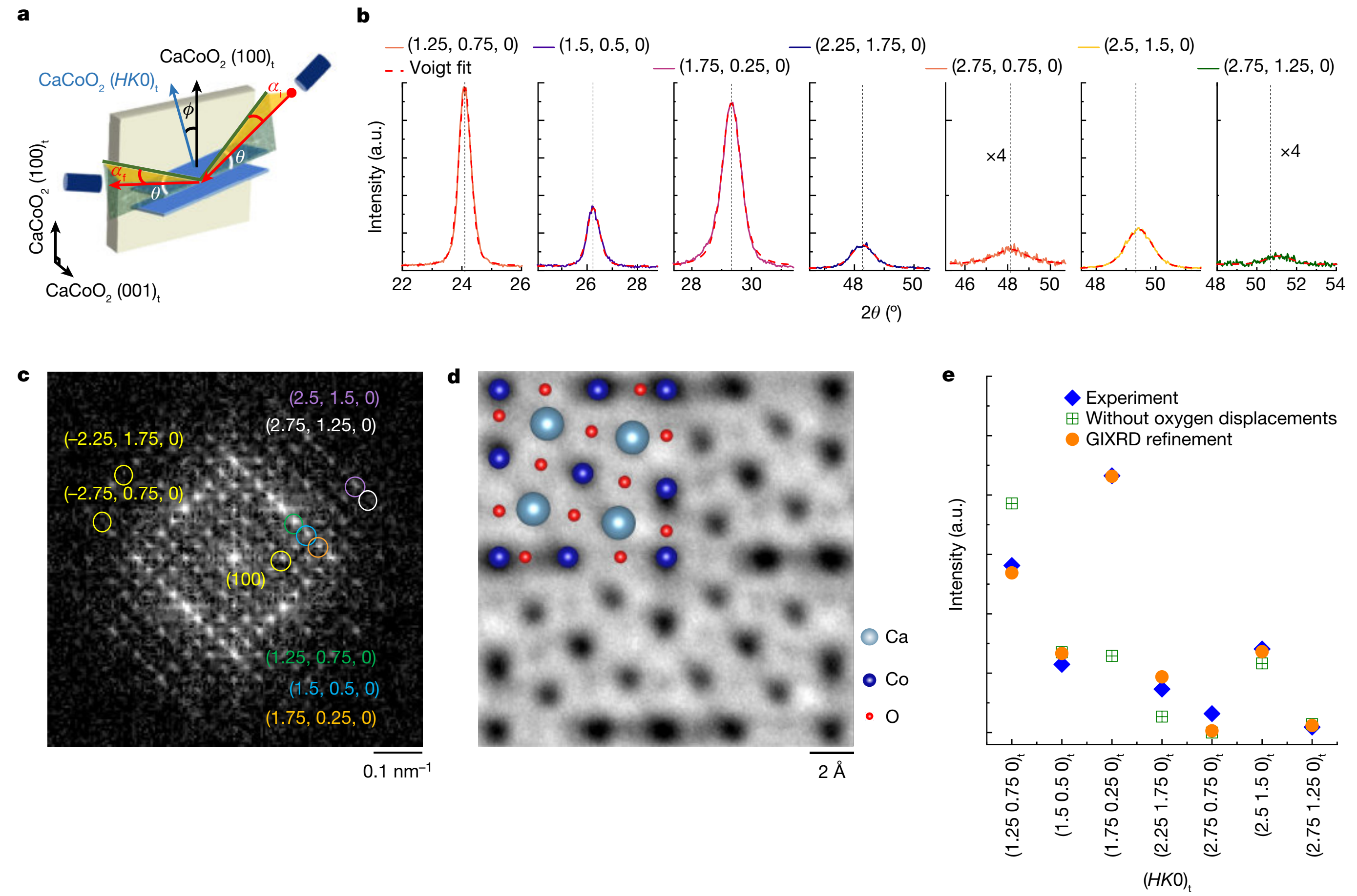


Fig. 3 | GIXRD and refinement of the oxygen positions. a, Schematic of the GIXRD experimental geometry. α_i and α_f are the incident and reflected angle, respectively, which are maintained to be 0.2° for all measurements. ϕ is the angle between $CaCoO_2$ $(100)_t$ and $(HK0)_t$ and θ is the diffraction angle for $CaCoO_2$ $(HK0)_t$ planes. **b**, GIXRD θ–2θ scans for various $CaCoO_2$ $(HK0)_t$ planes. The red dashed lines indicate Voigt fitting results. **c**, Fast Fourier transform of a plan-view HAADF-STEM image. All seven GIXRD peaks are observed. **d**, Unit-cell-averaged plan-view ABF-STEM image of $CaCoO_2$ with the final refined structure model overlayed. Dark contrast indicates the atomic columns. **e**, Relative integrated intensity for all $CaCoO_2$ $(HK0)_t$ GIXRD peaks. The blue (orange) filled diamonds (circles) are experimental (GIXRD refinement) results. The green crossed squares are calculated GIXRD intensity for $CaCoO_2$ in the absence of oxygen displacements (oxygen positions for an ideal square plane).

search for the actual scattering peaks, all of which were found very close to the simulated positions (Fig. 3b and Extended Data Table 2), confirming the overall structural symmetry deduced from the STEM results. The fast Fourier transform of the plan-view HAADF-STEM image in Fig. 2e also confirms the existence of the seven in-plane peaks observed from the GIXRD measurements (Fig. 3c). The approximate oxygen positions can also be extracted via plan-view annular bright-field (ABF)-STEM imaging (Fig. 3d). These were taken as the starting positions for Rietveld refinement of the relative intensities of the GIXRD peaks (blue diamonds in Fig. 3e; see Methods and Extended Data Tables 1 and 3). The refinement was dominated by oxygen displacements, such that the initial R factor of about 0.46 with undistorted oxygen (green squares in Fig. 3e) reduced to about 0.08 (orange circles in Fig. 3e), indicative of a high-quality fit. The overlay of the final refined structure of $CaCoO_2$ in Fig. 3d shows a good qualitative match with the experimental results within the precision of the ABF-STEM measurement.

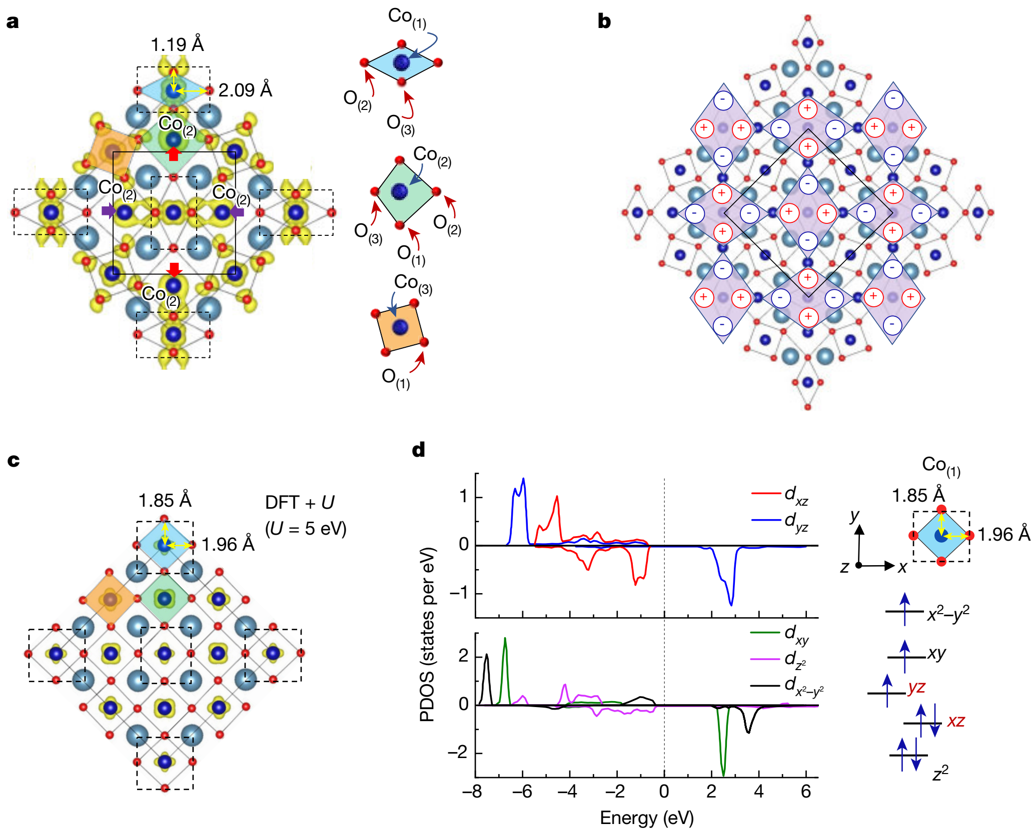
Origin of the ordered distortions in $CaCoO_2$

Figure 4a highlights the three distinct CoO_4 plaquettes (for $Co_{(1)}$, $Co_{(2)}$ and $Co_{(3)}$, coordinated by three distinct oxygen sites) in blue, green and orange, respectively, for the refined structure of $CaCoO_2$. The blue $Co_{(1)}O_4$ plaquette shows strong JT distortion as predicted[1]: the bonding length between $Co_{(1)}$ and $O_{(2)}$ ($d_{Co(1)-O(2)}$) is 2.09 ± 0.01 Å whereas that between $Co_{(1)}$ and $O_{(3)}$ ($d_{Co(1)-O(3)}$) is 1.19 ± 0.01 Å. It is noted that the anisotropic bonding ratio $d_{Co(1)-O(2)}/d_{Co(1)-O(3)}$ (1.76 ± 0.02) is extraordinarily large, for example, compared with the value for the JT-distorted octahedra in $LaMnO_3$ (about 1.15)[2]. The green $Co_{(2)}O_4$ plaquette is trapezoidal with three different Co–O bonding lengths, whereas that for $Co_{(3)}$ (orange) maintains four-fold symmetry (with rotation). Surprisingly, only 25% of the Co is strongly JT distorted, in contrast to the cooperative JTE in $LaMnO_3$ that provides an equivalent JT octahedral distortion for every Mn site (Fig. 1c).

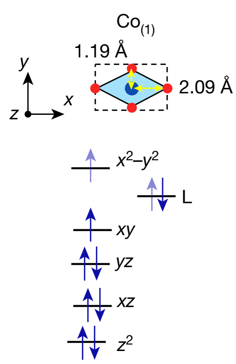
This intriguing situation can be understood to originate from the competition between the JT distortion and an induced geometrical frustration in the Ca layer. As shown by the dashed rectangles in Fig. 4a, the strong JTE distorts not only the $Co_{(1)}O_4$ but also the Ca layer. This unusual coupling of Ca to the Q_2 JT distortion arises from the lack of apical oxygen, which gives rise to strong electrostatic coupling between the CoO_2 and Ca layers. The anisotropic distortions of the Ca layer are however highly geometrically frustrated between neighbouring sites (Fig. 1f). Consequently, a complex arrangement of displacements is observed.

Interestingly, the geometrical frustration of the cooperative JTE results in a two-in–two-out type of distortion pattern, following ‘ice rules’[13] for the ångström-scale displacements of $Co_{(2)}$ around the maximally JT-distorted $Co_{(1)}O_4$ (Fig. 4a). The $Co_{(2)}O_4$ plaquette notably breaks

a



e



f

S (Degeneracy)	Non-JT $S = 3/2$ (8)	Small JT $S = 3/2$ (4)	Large JT $S = 1/2$ (2)
$3d_{x^2-y^2}$	0.91	0.92	0.92
$3d_{xy}$	1.00	1.00	1.00
$3d_{yz}$	0.50	1.00	1.00
$3d_{xz}$	0.50	0.00	0.00
$3d_{z^2}$	0.00	0.00	0.00
L	0.09	0.08	0.08

Fig. 4 | Extended structure of $CaCoO_2$, ice rules, quadrupolar ordering and electronic structure. a, Plan view of the experimentally refined crystal structure of $CaCoO_2$, consisting of three different CoO_4 sites corresponding to the three different colours (blue, green and orange). The blue-coloured CoO_4 plaquettes are strongly JT distorted. In the full unit cell, a pair of $Co_{(2)}$ ions distort towards (purple arrows) the central $Co_{(1)}$, whereas the other pair of $Co_{(2)}$ ions distort outwards (red arrows) from the central $Co_{(1)}$. The visualized isosurface in yellow represents the computed electron density for the lowest unoccupied $d_{x^2-y^2}$ band for $Co_{(1)}$ at the gamma point. **b**, The dipole arrangements associated with $Co_{(2)}$ displacements form an ordered array of electric quadrupoles (purple diamonds). **c**, Plan view of the computationally relaxed crystal structure of $CaCoO_2$ from DFT + U (U = 5 eV) calculations. This consists of three different CoO_4 sites, which show the same structural symmetry as experiment, but with much smaller distortions. The visualized isosurface in yellow colour represents the electron density for the lowest unoccupied $d_{x^2-y^2}$ band for $Co_{(1)}$ at the gamma point. **d**, The spin-dependent PDOS of $Co_{(1)}$ d orbitals from DFT + U and the corresponding atomic energy level diagram for the relaxed structure. **e**, An atomic energy level diagram appropriate for the experimentally refined structure from a ligand–multiplet calculation. The light blue arrows represent the occupation of half an electron. **f**, A table showing the number of holes from ligand–multiplet calculations. The three columns give results in the absence of distortions (non-JT), for parameters appropriate for the DFT + U relaxed structure (small JT) and for parameters appropriate for the experimentally refined structure (large JT).

inversion symmetry, and thus the long-range-ordered structure can be viewed as an ordering of electric dipoles. From an electrostatic point of view, this is equivalent to a 'herringbone' arrangement of electric quadrupoles (Fig. 4b), which minimizes the electrostatic interaction between quadrupole moments[32].

To further understand $CaCoO_2$, we performed density functional theory (DFT) + on-site Coulomb interaction (U) calculations. The overall symmetry for the relaxed structure directly matches the experimental observations, and it is robust to variations in U (Extended Data Fig. 5a). However, the magnitude of the distortions is substantially smaller—for example, the calculated anisotropic bonding ratio of about 1.06 is much less than the experimentally observed ratio of 1.76 ± 0.02. The electronic band structure of $CaCoO_2$ shows an insulating ground state with a bandgap energy of $E_g \approx 0.670$ eV (Extended Data Fig. 5b). Experimentally, the resistivity exhibits insulating temperature dependence (Extended Data Fig. 5c).

The spin-dependent partial density of states (PDOS) for $Co_{(1)}$ shows that the d_{z^2} and d_{xz} orbitals are doubly occupied (Fig. 4d) as expected

from strong JT splitting in a square-planar geometry (absence of apical ligands), and the d_{yz} orbital opens a gap between occupied and unoccupied states as indicated earlier (Fig. 1e). Overall, $Co_{(1)}$ is in a d^7 configuration where a spin-down hole resides in both the $d_{x^2-y^2}$ and d_{xy} orbitals, with an additional spin-down hole in d_{yz} orbitals (Figs. 1d and 4d). In contrast, the spin-dependent PDOS for $Co_{(3)}$ sites show degenerate d_{xz} and d_{yz} orbitals owing to the square symmetry of $Co_{(3)}O_4$ (Extended Data Fig. 5d,e). These results are consistent with expectations from the simple crystal field structure in Fig. 1e, indicating that the JTE, together with electron correlations, drives the symmetry-lowering distortion and lifts the orbital degeneracy.

We further compared the total energy of the non-JT-distorted structure, a $\sqrt{2}\times\sqrt{2}\times1$ supercell with a purely Q_2 JT distortion, and the relaxed structure to see whether the structure can be stabilized solely by Q_2 JT distortions. Interestingly, the result shows that $\sqrt{2}\times\sqrt{2}\times1$ supercell with purely Q_2 JT distortion is energetically more stable than a square-planar lattice without any distortion (Extended Data Fig. 6a). However, our DFT calculations with various starting points show that the structure always converged to the current DFT-relaxed structure, confirming that this structure minimizes the total energy (Extended Data Fig. 6b,c).

The quantitative discrepancy in the scale of distortions between experiment and theory may indicate the presence of large molecular orbital effects in this system that are difficult to capture using one-electron approaches[33,34]. Unlike octahedrally coordinated d^4 systems, the Q_2 JT distortion, which splits the d_{xz} and d_{yz} orbitals, draws the ligands along one direction inwards, towards $Co_{(1)}$. This substantially increases the hybridization of the σ-bonded $Co_{(1)}$–O orbitals (Fig. 4a), encouraging a strong admixture of holes on the neighbouring oxygen. Evidence for this tendency can be seen in the electron density distribution computed using the experimentally refined atomic positions (yellow isosurface in Fig. 4a), which is much more extended with substantial weight on oxygen when compared with the computationally relaxed structure (yellow isosurface in Fig. 4c). Thus a 'negative charge transfer' material may result, with trimerized molecular orbitals of Co ($3d_{x^2-y^2}$)–O ($2p_x$) hybrids, transitioning from high-spin Co $3d^7$ to a low-spin state Co $3d^8$L, where L denotes a ligand hole dispersed among the neighbouring σ-bonded oxygens (Fig. 4e). In fact, this could be the driving force behind the observed quadrupolar ordering (Fig. 4b).

Performing a small cluster exact diagonalization calculation for an 11-orbital unit cell (5 Co $3d$ and 3 O $2p_{x,y}$ orbitals; see Methods; result shown in Fig. 4f) confirms this picture. While a small JT splitting gives the DFT-relaxed configuration (high-spin $3d^7$), increasing the hybridization between Co and O along one direction, as in the experimental structure, draws electron weight from O $2p_y$ to Co $3d_{x^2-y^2}$ (Fig. 4f), raising the $3d_{x^2-y^2}$ complex above the $3d_{xy/xz}$ orbitals, further reducing the degeneracy. This low-spin $3d^8$L configuration is common to other Co-based negative-charge-transfer systems such as $SrCoO_{2.5}$ (ref. [35]) and $SrCoO_3$ (ref. [36]). Its presence in this system is further supported by O K-edge EELS (Extended Data Fig. 2d), which shows a pre-peak in good agreement with the hybridization pre-peak observed in $SrCoO_{3-\delta}$ ($\delta \le 0.5$)[37].

Conclusion

In summary, we have found $CaCoO_2$ to be a model system for realizing the 2D JT lattice with square-planar coordination. In contrast to layered perovskite-based systems with equivalent octahedral distortions[38–40], the interlayer coupling between the Ca layer and the CoO_2 layer is strong, such that the distorted structure is governed by the underlying competition between the JTE and the geometric frustration induced in the Ca layer. This leads to a complex pattern of distortions that arise in a $2\sqrt{2}a_t\times2\sqrt{2}a_t\times c_t$ unit cell and substantial ligand–transition metal mixing. The understanding of the JTE here may have general implications for complex compounds with 2D atomic layers where both interlayer coupling and the JTE are prominent.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-05681-2.

1. Jahn, H. A. & Teller, E. Stability of polyatomic molecules in degenerate electronic states I—orbital degeneracy. *Proc. R. Soc. Lond.* **161**, 220–235 (1937).
2. Goodenough, J. B. Theory of the role of covalence in the perovskite-type manganites [La,M(II)]MnO_3. *Phys. Rev.* **100**, 564–573 (1955).
3. Gehring, G. A. & Gehring, K. A. Co-operative Jahn–Teller effects. *Rep. Prog. Phys.* **38**, 1–89 (1975).
4. Goodenough, J. B. Jahn–Teller phenomena in solids. *Annu. Rev. Mater. Sci.* **28**, 1–27 (1998).
5. Goodenough, J. B. Jahn–Teller distortions induced by tetrahedral-site Fe^{2+} ions. *J. Phys. Chem. Solids* **25**, 151–160 (1964).
6. Siegrist, T., Zahurak, S. M., Murphy, D. W. & Roth, R. S. The parent structure of the layered high-temperature superconductors. *Nature* **334**, 231–232 (1988).
7. Smith, M. G., Manthiram, A., Zhou, J., Goodenough, J. B. & Markert, J. T. Electron-doped superconductivity at 40 K in the infinite-layer compound $Sr_{1-y}Nd_yCuO_2$. *Nature* **351**, 549–551 (1991).
8. Crespin, M., Levitz, P. & Gatineau, L. Reduced forms of $LaNiO_3$ perovskite. Part 1.—Evidence for new phases: $La_2Ni_2O_5$ and $LaNiO_2$. *J. Chem. Soc. Faraday Trans. 2* **79**, 1181–1194 (1983).
9. Hayward, M. A., Green, M. A., Rosseinsky, M. J. & Sloan, J. Sodium hydride as a powerful reducing agent for topotactic oxide deintercalation: synthesis and characterization of the nickel(I) oxide $LaNiO_2$. *J. Am. Chem. Soc.* **121**, 8843–8854 (1999).
10. Tsujimoto, Y. et al. Infinite-layer iron oxide with a square-planar coordination. *Nature* **450**, 1062–1065 (2007).
11. Kawakami, T. et al. Spin transition in a four-coordinate iron oxide. *Nat. Chem.* **1**, 371–376 (2009).
12. Dixon, E., Hadermann, J., Ramos, S., Goodwin, A. L. & Hayward, M. A. Mn(I) in an extended oxide: the synthesis and characterization of $La_{1-x}Ca_xMnO_{2+\delta}$ ($0.6 \le x \le 1$). *J. Am. Chem. Soc.* **133**, 18397–18405 (2011).
13. Pauling, L. The structure and entropy of ice and of other crystals with some randomness of atomic arrangement. *J. Am. Chem. Soc.* **57**, 2680–2684 (1935).
14. von Helmolt, R., Wecker, J., Holzapfel, B., Schultz, L. & Samwer, K. Giant negative magnetoresistance in perovskitelike $La_{2/3}Ba_{1/3}MnO_x$ ferromagnetic films. *Phys. Rev. Lett.* **71**, 2331–2333 (1993).
15. Raveau, B., Hervieu, M., Maignan, A. & Martin, C. The route to CMR manganites: what about charge ordering and phase separation? *J. Mater. Chem.* **11**, 29–36 (2001).
16. Han, J. E., Gunnarsson, O. & Crespi, V. H. Strong superconductivity with local Jahn–Teller phonons in C_{60} solids. *Phys. Rev. Lett.* **90**, 167006 (2003).
17. Millis, A. J., Littlewood, P. B. & Shraiman, B. I. Double exchange alone does not explain the resistivity of $La_{1-x}Sr_xMnO_3$. *Phys. Rev. Lett.* **74**, 5144–5147 (1995).
18. Millis, A. J., Shraiman, B. I. & Mueller, R. Dynamic Jahn–Teller effect and colossal magnetoresistance in $La_{1-x}Sr_xMnO_3$. *Phys. Rev. Lett.* **77**, 175–178 (1996).
19. Röder, H., Zang, J. & Bishop, A. R. Lattice effects in the colossal-magnetoresistance manganites. *Phys. Rev. Lett.* **76**, 1356–1359 (1996).
20. Guzmán-Verri, G. G., Brierley, R. T. & Littlewood, P. B. Cooperative elastic fluctuations provide tuning of the metal–insulator transition. *Nature* **576**, 429–432 (2019).
21. Keller, H., Bussmann-Holder, A. & Müller, K. A. Jahn–Teller physics and high-T_c superconductivity. *Mater. Today* **11**, 38–46 (2008).
22. Wurzenberger, X., Piotrowski, H. & Klüfers, P. A stable molecular entity derived from rare iron(II) minerals: the square-planar high-spin-d^6 $Fe_{II}O_4$ chromophore. *Angew. Chem. Int. Ed.* **50**, 4974–4978 (2011).
23. Li, D. et al. Superconductivity in an infinite-layer nickelate. *Nature* **572**, 624–627 (2019).
24. Zeng, S. et al. Phase diagram and superconducting dome of infinite-layer $Nd_{1-x}Sr_xNiO_2$ thin films. *Phys. Rev. Lett.* **125**, 147003 (2020).
25. Boullay, P. et al. Structure determination of a brownmillerite $Ca_2Co_2O_5$ thin film by precession electron diffraction. *Phys. Rev. B* **79**, 184108 (2009).
26. Xiang, L. et al. Exceptional oxygen evolution reactivities on $CaCoO_3$ and $SrCoO_3$. *Sci. Adv.* **5**, eaav6262 (2022).
27. Wang, Z. L. & Yin, J. S. Cobalt valence and crystal structure of $La_{0.5}Sr_{0.5}CoO_{2.25}$. *Phil. Mag. B* **77**, 49–65 (1998).
28. Li, H.-B. et al. Dehydration of electrochemically protonated oxide: $SrCoO_2$ with square spin tubes. *J. Am. Chem. Soc.* **143**, 17517–17525 (2021).
29. Hayward, M. A. et al. The hydride anion in an extended transition metal oxide array: $LaSrCoO_3H_{0.7}$. *Science* **295**, 1882–1884 (2002).
30. Lu, N. et al. Electric-field control of tri-state phase transformation with a selective dual-ion switch. *Nature* **546**, 124–128 (2017).
31. Osada, M. et al. Nickelate superconductivity without rare-earth magnetism: (La,Sr)NiO_2. *Adv. Mater.* **33**, 2104083 (2021).
32. Williams, J. H. The molecular electric quadrupole moment and solid-state architecture. *Acc. Chem. Res.* **26**, 593–598 (1993).
33. Radaelli, P. G. et al. Formation of isomorphic Ir^{3+} and Ir^{4+} octamers and spin dimerization in the spinel $CuIr_2S_4$. *Nature* **416**, 155–158 (2002).

34. Khomskii, D. I. & Streltsov, S. V. Orbital effects in solids: basics, recent progress, and opportunities. *Chem. Rev.* **121**, 2992–3030 (2021).
35. Chowdhury, S. et al. Negative charge-transfer energy in $SrCoO_{2.5}$ thin films: an interplay between O-2*p* hole density, charge-transfer energy, charge disproportionation, and ferromagnetic ordering. *ACS Appl. Electron. Mater.* **2**, 3859–3870 (2020).
36. Potze, R. H., Sawatzky, G. A. & Abbate, M. Possibility for an intermediate-spin ground state in the charge-transfer material $SrCoO_3$. *Phys. Rev. B* **51**, 11501–11506 (1995).
37. Cui, B. et al. Direct imaging of structural changes induced by ionic liquid gating leading to engineered three-dimensional meso-structures. *Nat. Commun.* **9**, 3055 (2018).
38. Hidaka, M., Inoue, K., Yamada, I. & Walker, P. J. X-ray diffraction study of the crystal structures of K_2CuF_4 and $K_2Cu_xZn_{1-x}F_4$. *Physica B+C* **121**, 343–350 (1983).
39. Aguado, F., Rodríguez, F., Valiente, R., Señas, A. & Goncharenko, I. Three-dimensional magnetic ordering in the Rb_2CuCl_4layer perovskite—structural correlations. *J. Phys. Condens. Matter* **16**, 1927–1938 (2004).
40. Cammarata, A. & Rondinelli, J. M. Ferroelectricity from coupled cooperative Jahn–Teller distortions and octahedral rotations in ordered Ruddlesden–Popper manganates. *Phys. Rev. B* **92**, 14102 (2015).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2023

Methods

Synthesis of $CaCoO_2$ thin films

Brownmillerite $CaCoO_{2.5}$ thin films were synthesized by pulsed laser deposition on perovskite $SrTiO_3$ (001) substrates. We used a $CaCoO_{2.5}$ polycrystalline target and a krypton fluoride excimer laser (wavelength $\lambda = 248$ nm, pulse repetition rate of 3 Hz and laser fluence of 1 J cm^{-2}). The distance between the target and the substrate was maintained at about 50 mm. The optimal conditions for high-quality $CaCoO_{2.5}$ films were found to be a substrate temperature $T = 600$ °C under oxygen partial pressure $P_{O_2} = 50$ mtorr. To protect the cobaltate films from potential degradation during and after the reduction process, we capped the as-grown $CaCoO_{2.5}$ with five-unit cells of $SrTiO_3$. The precursor $CaCoO_{2.5}$ film was then reduced to $CaCoO_2$ by topotactic reduction. We loosely covered the $CaCoO_{2.5}$ film with aluminium foil and sealed it with about 0.1 g CaH_2 powder in a vacuum glass tube. We annealed the sealed tube at 250 °C for 3 h in a tube furnace. The thermal ramping and cooling rates were 10 °C min^{-1}.

STEM measurements

STEM specimens were prepared using a Thermo Fisher Scientific Helios G4 UX focused ion beam. Cross-section samples were prepared and thinned to electron transparency using standard lift-out and thinning methods. Plan-view samples were prepared first using the standard lift-out method and then attached to a half grid at a 90° angle to the standard orientation and thinned. Total air exposure of the prepared STEM specimens was limited to less than 15 min to minimize possible degradation of the films.

HAADF and ABF-STEM data were acquired using a Thermo Fisher Scientific Spectra 300 X-CFEG operating at 120 kV or 300 kV with a convergence angle of 30 mrad. Inner and outer collection angles were approximately 60 mrad and 200 mrad, respectively, for HAADF and 15 mrad and 30 mrad, respectively, for ABF imaging. To obtain high-signal-to-noise ratio (SNR) images, many fast-acquisition images were acquired and subsequently aligned using a rigid registration process optimized for noisy images[41]. Owing to electron beam sensitivity of the sample, a probe current of less than 30 pA was used for all STEM data acquisitions. EELS measurements therefore required total acquisition times of about 5,000 s for each spectrum to achieve a sufficient SNR to distinguish differences in the EELS fine structure. The spectra presented are aligned and summed series of several acquisitions. To ensure stage stability sufficient for such long acquisitions, a Nion UltraSTEM operated at 100 kV and equipped with a high-stability stage, an Enfinium ER spectrometer and Quefina2 camera was used to acquire all EELS data. The effective energy resolution, measured by the full-width at half-maximum of the zero-loss peak, was about 0.34 eV. To ensure consistency, spectra from multiple regions of each sample were acquired. The individual spectra for each edge were first aligned with respect to each other and summed to produce high-SNR measurements for the as-grown and reduced films. Absolute energy alignments were then applied to the Co $L_{3,2}$ and Ca $L_{3,2}$ edges and the O K-edges based on the Ti $L_{3,2}$ and O K-edges of the $SrTiO_3$ substrate, respectively, which serve as well studied and consistent references. White line ratios were calculated from the Co $L_{3,2}$ edges using the Pearson method.

Synchrotron GIXRD

The synchrotron GIXRD experiments were performed at room temperature (293 K) using the Huber six-circle diffractometer at the 3A beamline of the Pohang Light Source-II in South Korea. The incident beam from an undulator source was monochromatized to 1.1078 Å. The incident beam angle was fixed to 0.2°. The synchrotron XRD data were collected in a 2θ range from 0° to 75° with a step interval of 0.01°.

Refinement of the oxygen positions

We developed an initial estimate for the oxygen positions based on ABF-STEM images and the available phonon distortion modes given the observed symmetry. We identified these phonon modes using the software package ISODISTORT, part of the ISOTROPY suite, which selects atomic displacement modes available when transitioning from a parent symmetry to a lower-symmetry structure[42,43]. We then used the application ISOVIZ to visualize these displacement modes and to manually identify the combination of modes that best matched the ABF-STEM contrast[42]. These initial oxygen positions (as well as the cation positions deduced from HAADF-STEM) were then used to further refine the GIXRD peaks by the Rietveld method. We used the least-squares method to minimize the function $M = \sum_i \{I_i^{obs} - I_i^{calc}\}^2$, where I_i^{obs} and I_i^{calc} are the observed and calculated integrated Bragg peak intensity of the ith (hkl) plane. We used seven independent (hkl) planes for the refinement. The calculated Bragg peak intensity for the (hkl) plane is given as $I_{hkl}^{calc} = Kp_{hkl}L_\theta\ |F_{hkl}|^2$, where K, p_{hkl}, L_θ and F_{hkl} are the scaling factor, number of equivalent (hkl) planes, Lorentz polarization factor and structure form factor, respectively. We used the R factor, $R = \frac{\sum_i |I_i^{obs} - I_i^{calc}|}{\sum_i I_i^{obs}}$, to quantify the goodness of fit.

Sample characterization

The conventional XRD data were taken using a monochromatized Cu Kα_1 source ($\lambda = 1.5406$ Å). The resistivity was measured using aluminium-wire-bonded contacts in a Hall bar geometry.

TOF-SIMS

TOF-SIMS characterization was performed using TOF.SIMS.5 NSC instrument (ION.TOF). A primary Bi^+ ion beam with an energy of 30 keV, d.c. current of 30 nA and a spot size of about 5 μm was used for extraction of the secondary ions of the analysed sample. A complimentary Cs^+ sputtering beam with an energy of 500 eV and a current of 40 nA was used for depth profiling. A low-energy unfocused electron flood gun was used for charge compensation. Measurements were performed in non-interlaced mode, where each scan with Bi^+ (100 μm × 100 μm, 4 s duration) was followed by a sputtering cycle with Cs^+ (300 μm × 300 μm, 2 s duration). A TOF mass analyser operated in positive-ion mode was used to analyse secondary ions with a mass resolution $m/\Delta m = 2{,}000$–$5{,}000$. The acquired chemical data provide one-dimensional depth profiles, characterizing the distribution of the elements of interest through the sample thickness. Depth calibration was performed using the known sample thickness in the assumption of homogeneous sputter rates.

DFT calculations

The initial structure was taken with all the atoms positioned at high-symmetry points of the simple tetragonal structure with a $2\sqrt{2}a_t \times 2\sqrt{2}a_t \times 2c_t$ supercell. Then structural relaxation was implemented with this initial structure using DFT + U calculations, where the Perdew–Burke–Ernzerhof[44] exchange-correlation functional was used and DFT + U was treated using the simplified (rotationally invariant) approach introduced by ref. [45]. All DFT + U calculations were performed using the Vienna Ab initio Simulation Package (VASP)[46]. The structural relaxation and self-consistent calculations were performed with 8 × 8 × 16 Monkhorst–Pack momentum points, whereas the density of states calculations were performed with 20 × 20 × 40 Monkhorst–Pack momentum points. It is noted that DFT + U calculations with various U values were explored, and 5 eV was chosen as the onset of opening a gap. The obtained relaxed structure shows the same structural symmetry breaking as found in the experimental structure but with smaller distortions. The magnitude and symmetry of the distortion are robust to variations in U values (Extended Data Fig. 5). To further investigate the stability of the distorted structure, we implemented generalized gradient approximation (GGA) + U, strongly constrained and appropriately normalized semilocal density functional (SCAN)[47], and SCAN + U calculations, on a series of structures with different distortions from zero up to the experimental distortion, as shown in Extended Data Fig. 6. We find that the DFT + U relaxed structure always shows

the lowest energy regardless of the calculation method used. Although none of the three methods predicts the amplitude of the experimental structure with the lowest energy, the symmetry-distorted structure (DFT + U relaxed structure) exhibits lower energy than the non-distorted one, showing the tendency of spontaneous symmetry breaking for the infinite-layer $CaCoO_2$ structure.

Cluster multiplet calculations

The calculations were performed by exact diagonalization of the CoO_2 11-orbital cluster involving 5 Co 3d orbitals and 6 O 2p orbitals forming 90° bond angles. The Hubbard–Kanamori interaction for O and Co orbitals captures multiplet energy levels parameterized by Slater–Condon integrals F. Hybridization is included through tight-binding hopping; charge transfer energy Δ and crystal field levels ε are chosen to give a high-spin Co $3d^7$ ground state (total spin 3/2, 8-fold ground-state degeneracy) for the undistorted plaquette. A JT perturbation that lifts the ground-state degeneracy is applied by changing relative orbital energies for $3d_{xz}$ and $3d_{yz}$, which maintains the high-spin ground state, while reducing the degeneracy to 4. Increasing the hybridization (hopping) along the x bond directions reduces the degeneracy to 2 and creates a low-spin ground state characterized by $3d^8$L with a ligand hole occupancy in the $2p_x$ orbital. Parameters are given in Extended Data Table 4.

Data availability

The data presented in the figures and other findings of this study are available from the corresponding authors upon reasonable request.

41. Savitzky, B. H. et al. Image registration of low signal-to-noise cryo-STEM data. *Ultramicroscopy* **191**, 56–65 (2018).
42. Campbell, B. J., Stokes, H. T., Tanner, D. E. & Hatch, D. M. ISODISPLACE: a web-based tool for exploring structural distortions. *J. Appl. Crystallogr.* **39**, 607–614 (2006).
43. Stokes, H. T., Hatch, D. M. & Wells, J. D. Group-theoretical methods for obtaining distortions in crystals: applications to vibrational modes and phase transitions. *Phys. Rev. B* **43**, 11010–11018 (1991).
44. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
45. Dudarev, S. L., Botton, G. A., Savrasov, S. Y., Humphreys, C. J. & Sutton, A. P. Electron-energy-loss spectra and the structural stability of nickel oxide: an LSDA+*U* study. *Phys. Rev. B* **57**, 1505–1509 (1998).
46. Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558–561 (1993).
47. Sun, J., Ruzsinszky, A. & Perdew, J. P. Strongly constrained and appropriately normed semilocal density functional. *Phys. Rev. Lett.* **115**, 036402 (2015).
48. Tassel, C. et al. $CaFeO_2$: a new type of layered structure with iron in a distorted square planar coordination. *J. Am. Chem. Soc.* **131**, 221–229 (2009).
49. Goodge, B. H. et al. Doping evolution of the Mott–Hubbard landscape in infinite-layer nickelates. *Proc. Natl Acad. Sci. USA* **118**, e2007683118 (2021).
50. Momma, K. & Izumi, F. VESTA: a three-dimensional visualization system for electronic and structural analysis. *J. Appl. Crystallogr.* **41**, 653–658 (2008).

Acknowledgements We thank W.-S. Lee for discussions. The work at SLAC and Stanford was supported by the US Department of Energy (DOE), Office of Basic Energy Sciences, Division of Materials Sciences and Engineering (contract number DE-AC02-76SF00515) and the Gordon and Betty Moore Foundation's Emergent Phenomena in Quantum Systems Initiative (grant number GBMF9072, synthesis equipment and initial development). Electron microscopy at Cornell was support by the Department of Defense Air Force Office of Scientific Research (number FA 9550-16-1-0305) and the Packard Foundation, and made use of the Cornell Center for Materials Research Shared Facilities which are supported through the NSF MRSEC programme (DMR-1719875), with the Thermo Fisher Helios G4 UX focused ion beam also supported by NSF (DMR-1539918). The Thermo Fisher Spectra 300 X-CFEG was acquired with support from PARADIM, an NSF MIP (DMR-2039380), and Cornell University. M.A.S. acknowledges additional support from the NSF GRFP under award number DGE-1650441. The 3A beamline at PLS-II is supported in part by MSIT. B.-G.C. is currently affiliated to Korea Research Institute of Standards and Science (KRISS). D.J. acknowledges funding by the Alexander-von-Humboldt foundation via a Feodor Lynen postdoctoral fellowship. Raman spectroscopy measurement was performed at the Stanford Nano Shared Facilities (SNSF), supported by the National Science Foundation under award ECCS-2026822. TOF-SIMS characterization was conducted at the Center for Nanophase Materials Sciences, which is a DOE Office of Science User Facility, and using instrumentation within ORNL's Materials Characterization Core provided by UT-Battelle, LLC under contract number DE-AC05-00OR22725. The computational work for this project was performed on the Sherlock cluster in the Stanford Research Computing Center.

Author contributions W.J.K. and H.Y.H. conceived and designed the experiments. M.A.S., B.H.G. and L.F.K. performed the STEM and EELS measurements and analysis. C.J. performed the DFT calculations. C.J., B.M. and T.P.D. performed the cluster calculations. D.J. performed Raman spectroscopy measurements. W.J.K. grew the samples, which were characterized by W.J.K., K.L., D.J. and M.O. W.J.K. and B.-G.C. performed and analysed the synchrotron GIXRD measurements. A.V.I. performed TOF-SIMS measurements. W.J.K., T.P.D. and H.Y.H. wrote the manuscript, with input from all authors.

Competing interests The authors declare no competing interests.

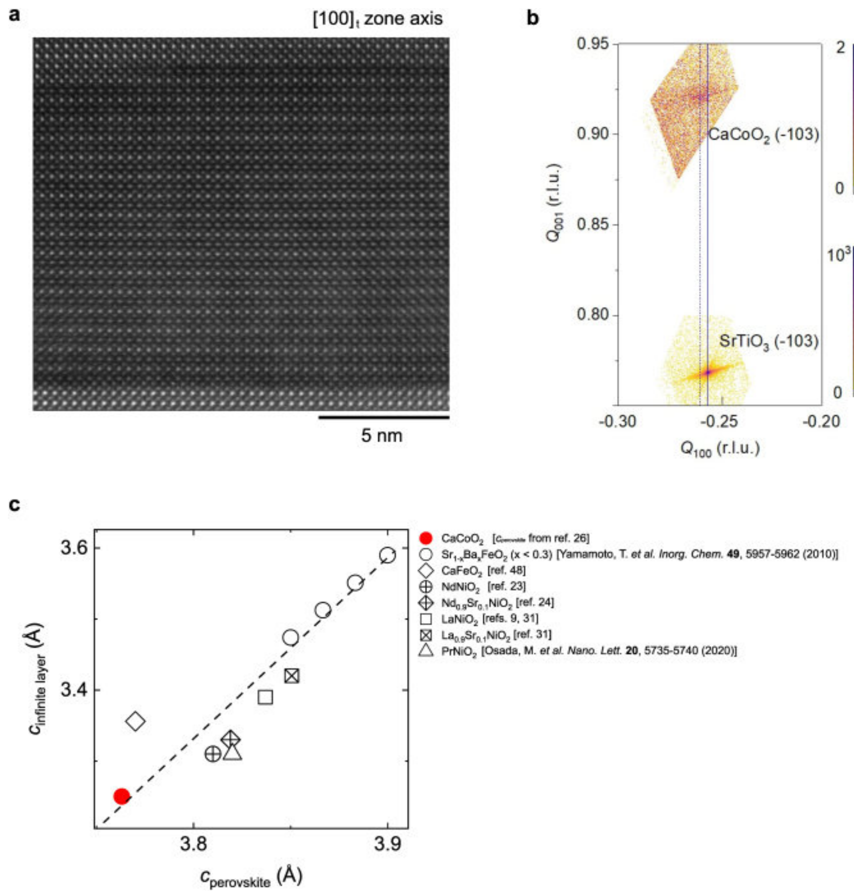
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-05681-2.

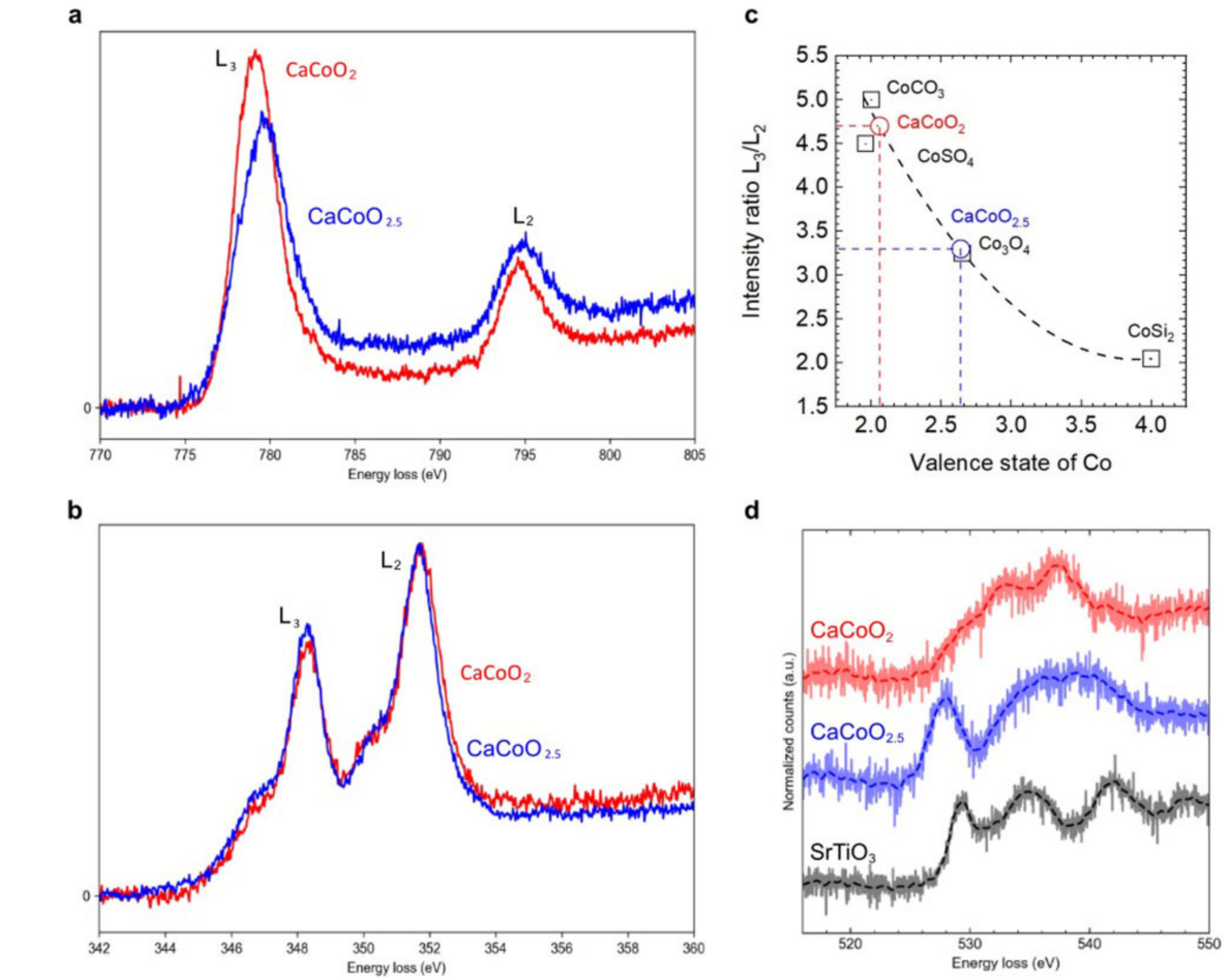
Correspondence and requests for materials should be addressed to Woo Jin Kim or Harold Y. Hwang.

Peer review information *Nature* thanks Eric Hoglund and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

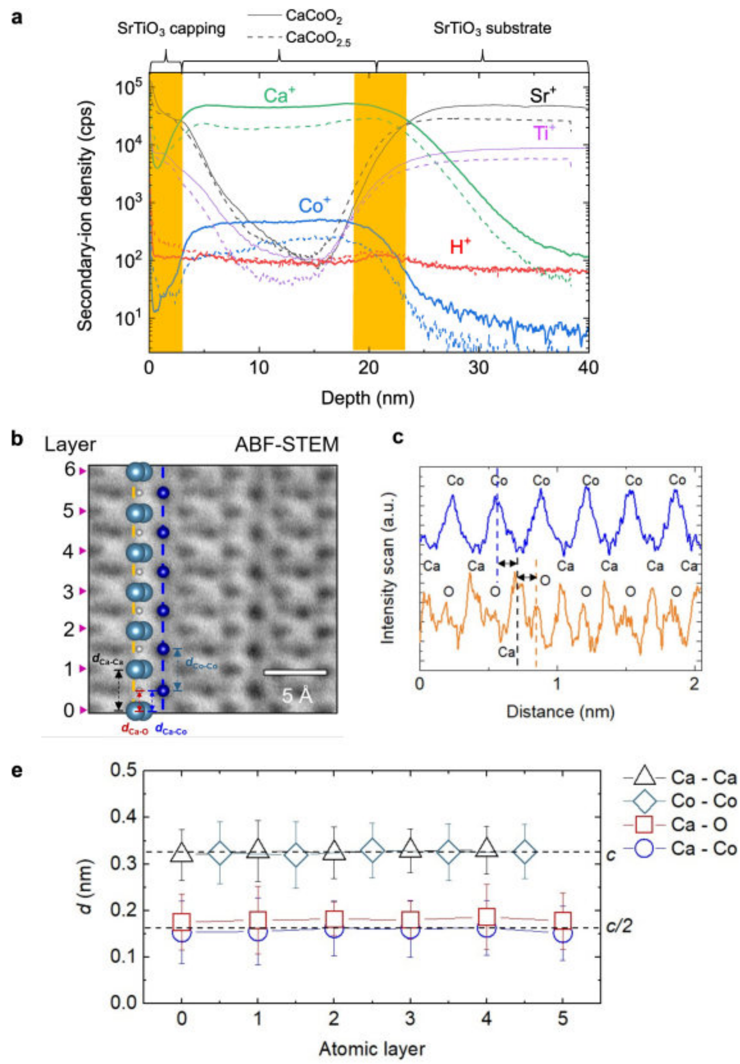
Reprints and permissions information is available at http://www.nature.com/reprints.



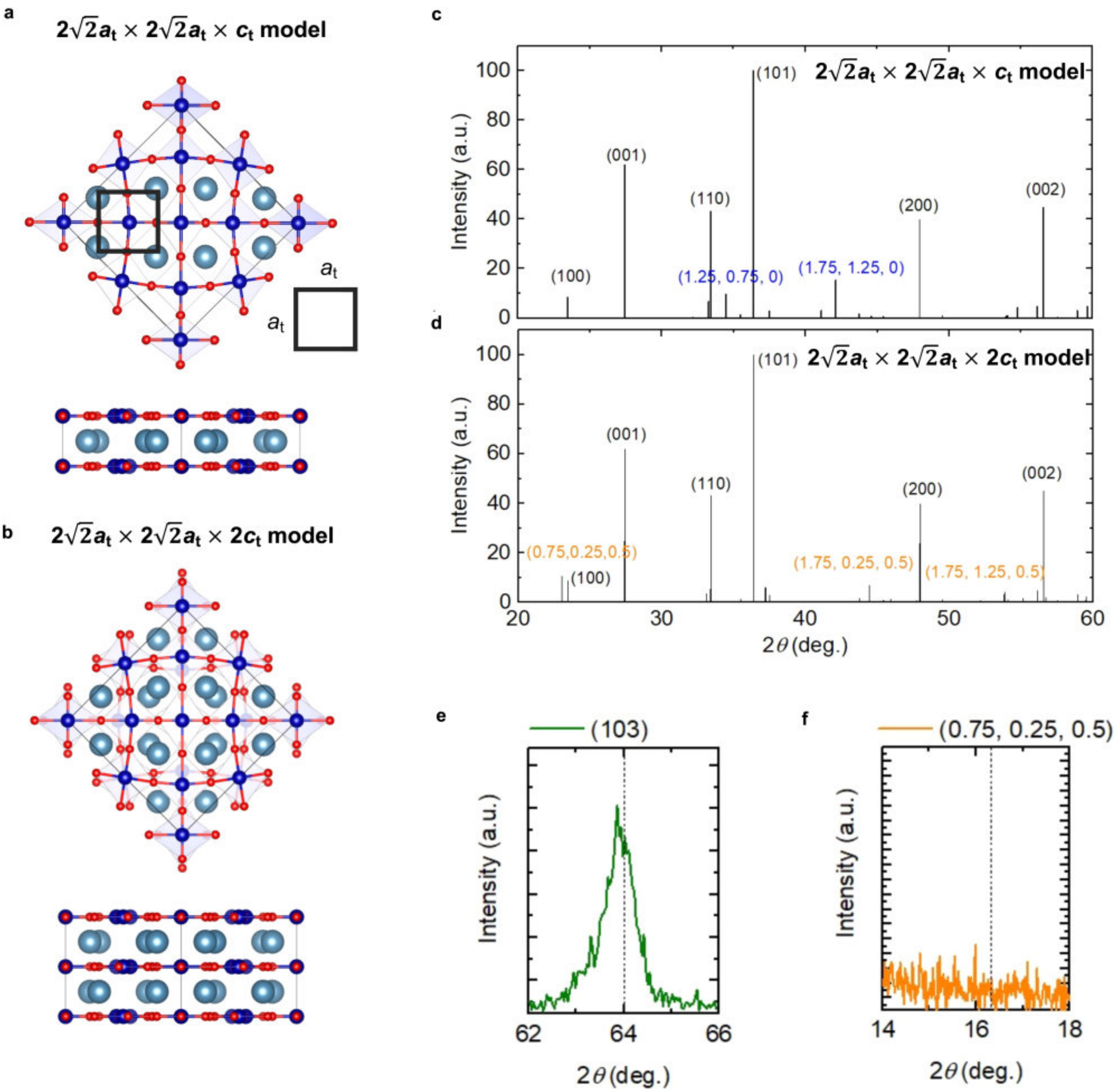
Extended Data Fig. 1 | Structural characterizations for $CaCoO_{2.5}$ and $CaCoO_2$. a, Atomic-resolution HAADF-STEM image along the $[100]_t$ zone-axis projection of $CaCoO_{2.5}$ showing alternate stacking of tetrahedral and octahedral layers. **b**, X-ray diffraction reciprocal space map of $CaCoO_2$ around the (−103) $SrTiO_3$ diffraction peak, indicating that the film is relaxed from the substrate. **c**, Empirical relationship between perovskite and infinite-layer lattice parameters. *c*-axis lattice parameters for various transition metal oxide compounds are plotted for the perovskite phase and the infinite-layer phase after topotactic reduction. The dashed line is a linear fit for all the data points in the plot. Note that the $CaFeO_2$ has relatively large $c_{\text{infinite layer}}/c_{\text{perovskite}}$ associate with out-of-plane displacement of both FeO_4 and Ca layers[48].



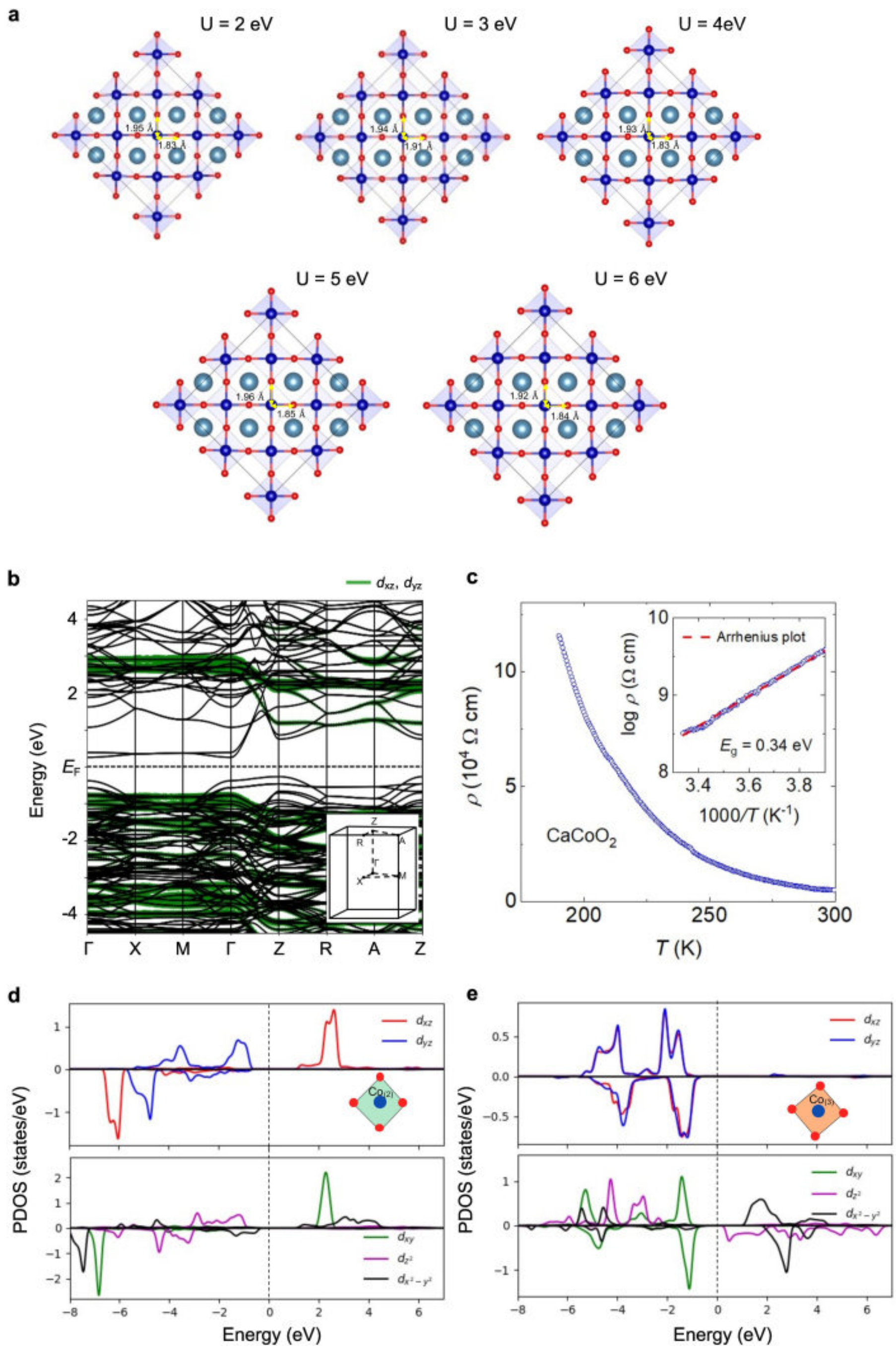
Extended Data Fig. 2 | EELS measurements of $CaCoO_2$. **a**, Co-$L_{3,2}$ edge; the blue (red) solid line indicates EEL spectra for $CaCoO_{2.5}$ ($CaCoO_2$). **b**, Ca-$L_{3,2}$ edge EELS shows that there are no substantial changes in the spectra before ($CaCoO_{2.5}$, blue) and after reduction ($CaCoO_2$, red). **c**, A plot of the intensity ratio $I(L_3)/I(L_2)$ of the Co-$L_{3,2}$ edge for different Co compounds with different oxidation states. Note that the dashed line indicates a polynomial fit curve for four different compounds from ref. [27] ($CoCO_3$, $CoSO_4$, Co_3O_4, and $CoSi_4$). $I(L_3)/I(L_2)$ of the $CaCoO_{2.5}$ and $CaCoO_2$ films are depicted with blue and red circles, respectively. **d**, O K-edges EELS data. Spatially averaged O K-edge spectra of $CaCoO_2$ ($CaCoO_{2.5}$) in red (blue). The partially transparent, solid lines indicate the raw, background-subtracted data, and the dashed lines indicate the Gaussian filtered spectra. Upon reduction of the $CaCoO_{2.5}$ films to $CaCoO_2$, we observe a suppression of the distinct pre-peak at ~ 529 eV in the region of the O K-edge associated with hybridization between O 2p and transition metal d orbitals consistent with a nominal electronic transition from $3d^6$ to $3d^7$. This is similar to the pre-peak suppression observed upon reduction from perovskite to infinite-layer phase in the related nickelates[49]. We further see the emergence of a shoulder in the $CaCoO_2$ spectrum at ~ 530 eV, which is similar to a feature attributed to ligand hole states in doped infinite-layer nickelates[49]. This feature is also consistent with published spectra acquired from $SrCoO_{3-\delta}$, which has negative charge transferred state[37]. An O K-edge spectrum of the $SrTiO_3$ substrate is included in black for comparison.



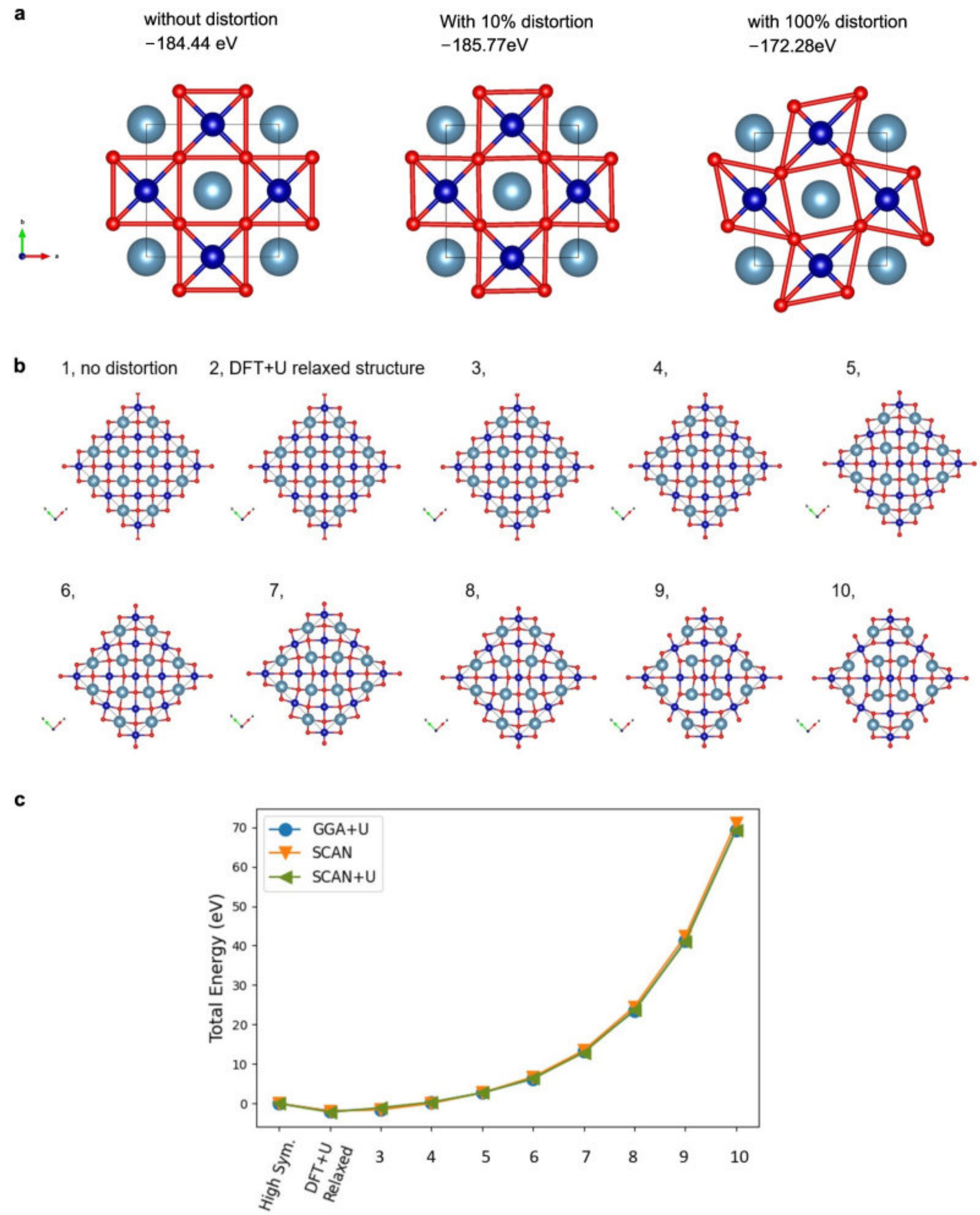
Extended Data Fig. 3 | Time-of-flight secondary-ion mass spectrometry (TOF-SIMS) and ABF-STEM measurements of $CaCoO_2$. a, Depth profiles of H^+ and other ions from both $CaCoO_{2.5}$ and $CaCoO_2$ thin film on $SrTiO_3$ substrate (with ~ 2 nm $SrTiO_3$ capping layer) were measured with secondary-ion mass spectrometry. The Co ion signals from both $CaCoO_{2.5}$ and $CaCoO_2$ thin films were employed as a marker for the interface position. TOF-SIMS measurements show that the H^+ concentration for $CaCoO_2$ is similar to the background level of the as-synthesized $CaCoO_{2.5}$ thin film. **b**, ABF-STEM image along the $[100]_t$ zone-axis projection with overlaid Co, Ca, and O atoms. **c**, Intensity line profiles for the blue and the orange dashed lines in **b**. The intensities of the line profiles are from inverted image **b**. The blue (orange) solid line indicates the line profile for the Co column (Ca and O column). The peak positions are the relative distances noted at the bottom of the image **b**. **d**, Atomic distances between Ca and Ca (black triangles), Co and Co (green diamonds), Ca and O (red squares), and Ca and Co (blue circles) layers are plotted. Note that the atomic layer numbers in **d** correspond to those in **b**. Error bars are taken as the full-width at half-maximum of the intensity peaks in **c**.



Extended Data Fig. 4 | Powder XRD simulation and *c*-lattice parameter determination. Lattice structure models for **a**, $2\sqrt{2}a_t \times 2\sqrt{2}a_t \times c_t$ and **b**, $2\sqrt{2}a_t \times 2\sqrt{2}a_t \times 2c_t$. The second structure model is lattice doubled from the first model by stacking a half-unit-cell shifted layer along the in-plane direction. Powder XRD simulation results for both **c**, $2\sqrt{2}a_t \times 2\sqrt{2}a_t \times c_t$ and **d**, $2\sqrt{2}a_t \times 2\sqrt{2}a_t \times 2c_t$ models. Note that the XRD simulation for the $2\sqrt{2}a_t \times 2\sqrt{2}a_t \times 2c_t$ model has a distinct half-order peak along the *c*-lattice direction. We first found **e**, the $CaCoO_2$ $(103)_t$ XRD peak as a reference peak. Based on this reference peak position, we perform θ–2θ scans along the expected $CaCoO_2$ (0.75, 0.25, 0.5) position. **f**, No XRD peak was observed at the expected $CaCoO_2$ (0.75, 0.25, 0.5) peak position, indicating that $CaCoO_2$ does not have a *c*-axis doubling of the simple tetragonal unit cell.



Extended Data Fig. 5 | DFT calculations for $CaCoO_2$. **a**, Plan-view of the relaxed crystal structure for $CaCoO_2$ from DFT + U calculations with U = 2 eV, U = 3 eV, U = 4 eV, U = 5 eV, and U = 6 eV. **b**, Calculated band dispersion of $CaCoO_2$ (DFT + U for U = 5 eV). Green highlights d_{xz} (and d_{yz}) projections. The inset shows high-symmetry points in the tetragonal Brillouin zone. **c**, Resistivity versus temperature of $CaCoO_2$ thin film. The inset shows that the resistivity is well fitted with an Arrhenius plot with an estimated (transport) gap of 0.337 ± 0.001 eV. The spin-dependent partial density of states (PDOS) of **d**, $Co_{(2)}$ and **e**, $Co_{(3)}$ d orbitals from DFT + U (U = 5 eV). The spin-dependent PDOS of $Co_{(2)}$ shows the degeneracy lifting of the $d_{xz/yx}$-orbitals.



Extended Data Fig. 6 | Total energy calculation for $CaCoO_2$ with $\sqrt{2} \times \sqrt{2} \times 1$ and $2\sqrt{2} \times 2\sqrt{2} \times 1$ supercell. a, DFT+U (U = 5 eV) calculations for the total energy under purely Q_2-JT-distortions in the $\sqrt{2} \times \sqrt{2} \times 1$ supercell. **b**, $2\sqrt{2} \times 2\sqrt{2} \times 1$ supercell with different distortion amplitudes. Approaching #10, the structure is approaches the experimentally refined structure. **c**, Normalized total energy for the structures depicted in **b**. Three different first-principle calculations are used for **c** (Methods).

Extended Data Table 1 | Atomic coordinates for the initial (before GIXRD refinement) and refined structure of $CaCoO_2$

P42₁2 symmetry group

Atom	x		y		z		Occupation
	Initial	Refined	Initial	Refined	Initial	Refined	
$Co_{(1)}$	0	0	0	0	0	0	1
$Co_{(2)}$	0.200	0.205 ± 0.011	0.200	0.205 ± 0.011	0	0	1
$Co_{(3)}$	0.500	0.500	0	0	0	0	1
Ca	0.460	0.459 ± 0.012	0.250	0.255 ± 0.011	0.500	0.500	1
$O_{(1)}$	0.349	0.349 ± 0.010	0.099	0.099 + 0.010	0	0	1
$O_{(2)}$	0.349	0.363 ± 0.012	0.349	0.363 ± 0.012	0	0	1
$O_{(3)}$	0.099	0.078 ± 0.011	0.099	0.078 ± 0.011	0	0	1

Initial atomic coordinates were taken from the values from the STEM measurements.

Extended Data Table 2 | Simulated and experimental $CaCoO_2$ $(HK0)_t$ GIXRD peak positions and intensities

Photon energy 11.2 keV (0.11078 nm)

h	*k*	*l*	$\phi_{sim.}$ (°)	$\phi_{exp.}$ (°)	$2\theta_{sim.}$ (°)	$2\theta_{exp.}$ (°)	Intensity$_{sim.}$ (a.u.)	Intensity$_{exp.}$ (a.u.)
1.25	0.75	0	30.96	30.83 ± 0.002	24.47	24.10 ± 0.001	9.92	1406 ± 13
0.5	1.5	0	18.43	18.31 ± 0.002	26.58	26.24 ± 0.002	3.03	574 ± 10
1.75	0.25	0	8.13	7.72 ± 0.002	29.79	29.33 ± 0.002	15.10	2162 ± 20
2.25	1.75	0	37.88	37.74 ± 0.003	48.96	48.24 ± 0.004	5.04	366 ± 8
2.75	0.75	0	15.26	15.64 ± 0.032	48.96	48.14 ± 0.027	0.90	160 ± 29
2.5	1.5	0	30.96	31.02 ± 0.003	50.16	49.43 ± 0.003	3.48	704 ± 10
2.75	1.25	0	24.44	24.53 ± 0.035	52.10	51.10 ± 0.034	0.36	46 ± 7

VESTA software[50] was used for the simulated GIXRD peak positions and intensities.

Extended Data Table 3 | Structure symmetry and atomic coordinates for the refined structure of $CaCoO_2$

Data collection	$CaCoO_2$ (GIXRD refinement)				
Space group	$P42_12$ (No. 90)				
Cell dimensions					
a, b, c (Å)	$a = b = 10.78$ Å, $c = 3.27$ Å				
a , b, g (°)	α = b = g= 90°				
Atomic coordination	*Atom*	*x*	*y*	*z*	*Occupation*
	$Co_{(1)}$	0	0	0	1
	$Co_{(2)}$	0.205 ± 0.011	0.205 ± 0.011	0	1
	$Co_{(3)}$	0.500	0	0	1
	Ca	0.459 ± 0.012	0.255 ± 0.011	0.500	1
	$O_{(1)}$	0.349 ± 0.010	0.099 ± 0.010	0	1
	$O_{(2)}$	0.363 ± 0.012	0.363 ± 0.012	0	1
	$O_{(3)}$	0.078 ± 0.011	0.078 ± 0.011	0	1

Extended Data Table 4 | Parameters used for multiplet calculations (in eV)

F_0	6
F_2	0.12857
F_4	0.025
Coulomb U_{dd}	7.4
Hund's J	0.886
Average Coulomb $\bar{U}$	5.45
$d_{x2\text{-}y2}$	0
d_{z2}	2.4
d_{xy}	0.2
$d_{xz/yz}$	1.2
Δ	$3\bar{U}$
t_{pd}	1
t_{pdz}	$t_{pd}/4$
$t_{pdxy} = t_{pdxz} = t_{pdyz}$	0.225
$t_{pp\sigma}$	0.25
Coulomb U_{pp}	1

These parameters were used to generate the ground-state ($3d^7$) configuration for the cluster. The high-spin Jahn–Teller split ground state was obtained by lifting the degeneracy of the $d_{xz/yz}$ site energies, while the low-spin ground state was obtained by increasing all hybridizations along the *x*-direction of the cluster.

Evidence of near-ambient superconductivity in a N-doped lutetium hydride

https://doi.org/10.1038/s41586-023-05742-0

Received: 26 April 2022

Accepted: 18 January 2023

Published online: 8 March 2023

Check for updates

Nathan Dasenbrock-Gammon[1,4], Elliot Snider[2,4], Raymond McBride[2,4], Hiranya Pasan[1,4], Dylan Durkee[1,4], Nugzari Khalvashi-Sutter[2], Sasanka Munasinghe[2], Sachith E. Dissanayake[2], Keith V. Lawler[3], Ashkan Salamat[3] & Ranga P. Dias[1,2]✉

The absence of electrical resistance exhibited by superconducting materials would have enormous potential for applications if it existed at ambient temperature and pressure conditions. Despite decades of intense research efforts, such a state has yet to be realized[1,2]. At ambient pressures, cuprates are the material class exhibiting superconductivity to the highest critical superconducting transition temperatures (T_c), up to about 133 K (refs. [3–5]). Over the past decade, high-pressure 'chemical precompression'[6,7] of hydrogen-dominant alloys has led the search for high-temperature superconductivity, with demonstrated T_c approaching the freezing point of water in binary hydrides at megabar pressures[8–13]. Ternary hydrogen-rich compounds, such as carbonaceous sulfur hydride, offer an even larger chemical space to potentially improve the properties of superconducting hydrides[14–21]. Here we report evidence of superconductivity on a nitrogen-doped lutetium hydride with a maximum T_c of 294 K at 10 kbar, that is, superconductivity at room temperature and near-ambient pressures. The compound was synthesized under high-pressure high-temperature conditions and then—after full recoverability—its material and superconducting properties were examined along compression pathways. These include temperature-dependent resistance with and without an applied magnetic field, the magnetization (M) versus magnetic field (H) curve, a.c. and d.c. magnetic susceptibility, as well as heat-capacity measurements. X-ray diffraction (XRD), energy-dispersive X-ray (EDX) and theoretical simulations provide some insight into the stoichiometry of the synthesized material. Nevertheless, further experiments and simulations are needed to determine the exact stoichiometry of hydrogen and nitrogen, and their respective atomistic positions, in a greater effort to further understand the superconducting state of the material.

Dense elemental hydrogen has long been predicted to be a very-high-temperature superconductor[22,23], yet the extremely high pressures required have presented challenges in confirming those superconducting phases[24,25]. The superhydride materials offer the promise of retaining the superconducting properties of dense elemental hydrogen but at much lower pressures. The prediction of a 220–235-K superconducting transition temperature (T_c) in CaH_6 at 150 GPa (ref. [26]) and the watershed discovery of a 203-K T_c for H_3S at 155 GPa (ref. [27]) have instigated a materials discovery boon in which, at present, almost all possible binary systems of high-pressure hydride systems have been modelled[28]. The recent observation of an anomalously high T_c in YH_6 showed that high-temperature superconductivity can be achieved with lower hydrogen content and more modest pressures than previously understood[13]. As the main discoveries have all been at greater than megabar pressures, the goal has shifted to further lowering the pressure required, with a focus on the vast sample space of ternary hydride compounds. One direction is a third, light element acting as a dopant in the metal hydrides, which is predicted to have two main beneficial effects[29]. First, there is a predicted increase in T_c as seen in proposed examples such as critical temperatures approaching 500 K in the Li–Mg–H system, although still in the megabar regime[17], and virtual crystal approximation simulations and recent experimental evidence indicating an increase of the transition temperature by at least 25 K from doping the LaH_{10} framework[18,19,30]. Second, the addition of a third element can greatly enhance the stability of a hydrogen-rich lattice, thereby lowering the pressure range over which it is stable. $LaBH_8$ is predicted to be stable down to 20–40 GPa while maintaining its high-temperature superconductivity[20,21], and a metal–boron–carbon clathrate is predicted to retain its superconducting properties at ambient pressures[31]. The presence of stability combined with increasing T_c by introducing the third element opens the possibility of pushing the hydride superconductors to higher values of T_c at sub-megabar pressures.

[1]Department of Physics and Astronomy, University of Rochester, Rochester, NY, USA. [2]Department of Mechanical Engineering, School of Engineering and Applied Sciences, University of Rochester, Rochester, NY, USA. [3]Unearthly Materials Inc., Rochester, NY, USA. [4]These authors contributed equally: Nathan Dasenbrock-Gammon, Elliot Snider, Raymond McBride, Hiranya Pasan, Dylan Durkee. ✉e-mail: rdias@rochester.edu

As there is an overwhelming amount of phase space unexplored by simulation in ternary rare-earth hydrides, rational chemical design is needed at present to identify the next candidate material. The La and Y binary superhydrides are predicted and measured to adopt similar high-pressure stoichiometries and phases with the Y-based ones exhibiting higher T_c at equivalent pressures[8–10,12]. The smaller size of the Y^{3+} cation offers a simple chemical rationale for this behaviour. However, the Sc hydrides with an even smaller ionic radius are predicted to have completely different structures and lower T_c (ref. [32]). Owing to the lanthanoid contraction, the lanthanoids heavier than Dy offer comparable or smaller trivalent ionic radii than Y but with the complication of *f* electrons[33–35]. Although the 4*f* electrons in lanthanoid compounds are often atom-localized and semivalent at ambient conditions, the inherent magnetism of partially occupied 4*f* states[36–38] or migration towards the Fermi level under pressure could be detrimental to the superconducting properties[9,39]. Although synthesis efforts for the high-pressure YbH_x system have produced structures distinct from its La and Y counterparts, probably owing to the transfer of *d* electrons to unoccupied *f* states[40], predictions indicate that hydrides of the two heaviest lanthanoids should be able to achieve $T_c \geq 145$ K by a megabar owing to the strong electron correlation of the 4*f* electrons near the Fermi level[41]. Causes of the high T_c achieved in the sub-megabar regime are believed to be twofold. First, the over half-filled valence 4*f* states suppress the phonon softening and second they provide some enhancement to the electron–phonon coupling relative to the transition metal (Y and La) rare earths. Combining the benefits of light atom doping and the presence of 4*f* electrons in the valance states should increase the stability of a hydrogen-rich rare-earth hydride to lower pressures while potentially enhancing its superconducting properties.

In this paper, we present experimental evidence of superconductivity at 294 K and 10 kbar pressure in a ternary lutetium–nitrogen–hydrogen compound in which the combination of a full 4*f* shell along with the electron donation and chemical pressure of the nitrogen drive the T_c and pressure stability of nitrogen-doped lutetium hydride into the near-ambient regime. The measured superconducting properties are the observation of zero resistance, a.c. magnetic susceptibility and d.c. magnetic susceptibility with zero field and field cooling, magnetization *M*–*H* curve, heat capacity, voltage–current (*V*–*I*) curves and the reduction of T_c under an external magnetic field with an upper critical magnetic field of about 88 tesla based on the Ginzburg–Landau (GL) model at zero temperature (see Extended Data Fig. 15). The composition and structure are explored with elemental analysis, EDX measurements, XRD, Raman spectroscopy and density functional theory (DFT) simulations.

Temperature–pressure relations and visual change

The near-ambient superconducting stability regime of the ternary lutetium–nitrogen–hydrogen system extends from about 3 kbar to about 30 kbar (Fig. 1a) and is accompanied by a marked visual transformation over just a few kbar of pressure (Fig. 1b). The recovered sample is initially in a non-superconducting metallic phase with a lustrous bluish colour, denoted here as phase I. Compression to about 3 kbar drives the progression of the system into phase II, leading to the onset of the superconducting regime, and this transformation is associated with a sudden change in colour from blue to pink. T_c as determined by electrical-resistance, magnetic-susceptibility and heat capacity-measurements increases with pressure from 171 K at around 5 kbar until it peaks at 294 K at around 10 kbar. The 10-kbar turning point in the superconducting dome is followed by a reduction in T_c to approximately 200 K before about 30 kbar. Compression above the highest pressure T_c shown in Fig. 1a drives the sample through another phase transition into phase III. Phase III is a non-superconducting metallic state that is once again distinct in colour, being bright red in appearance. The colour changes are for reflected light only and the sample is completely opaque for transmitted light.

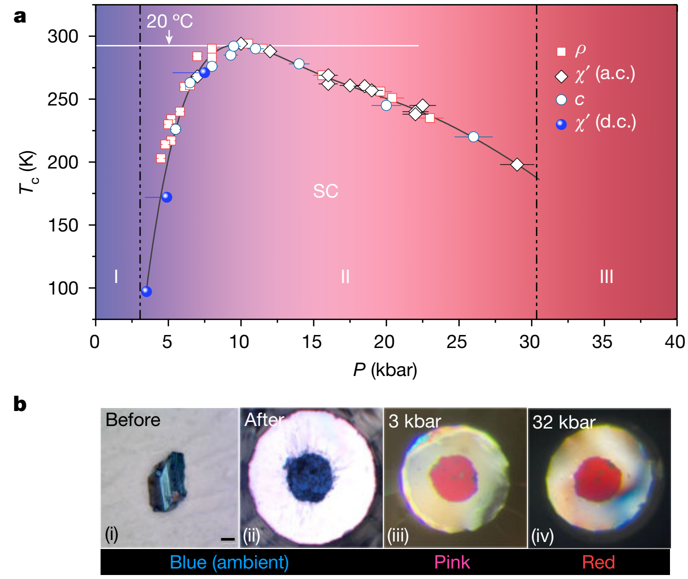


Fig. 1 | Superconductivity in lutetium–nitrogen–hydrogen at near-ambient pressures. a, Evolution of the superconducting transition temperature (T_c) of recompressed nitrogen-doped lutetium hydride as a function of pressure (*P*), illustrating a clear dome-shaped peak around 10 kbar with a T_c of 294 K. Superconductivity (SC) is only observed in phase II between about 3 kbar and about 30 kbar of pressure. *ρ*, electrical resistance; *χ*′ (a.c.), dynamic magnetic susceptibility; *χ*′ (d.c), static magnetic susceptibility; *c*, heat capacity. **b**, Microphotographs of a sample using only reflected light recovered from a metallic sample from high-pressure high-temperature synthesis before being loaded into the diamond anvil cell (i) and after loading at ambient pressure, showing a remarkable blue colour (ii). Scale bar, 10 µm. The sample is completely opaque to transmitted light. Under pressure, the sample transformed from blue to pink at about 3 kbar (iii) and to red at about 30 kbar (iv).

Temperature-dependent electrical resistance

The superconducting transitions of phase II are evidenced by a sharp drop in resistance within a temperature change of a few degrees kelvin (Fig. 2a). The temperature-dependent resistance data were acquired during the natural warming cycle (about 0.25 K min^{-1}) with a current of 100 µA to 2 mA. The accuracy of the temperature probe is ±0.1 K and the transition temperature was determined from the onset of superconductivity. In all experiments, the reported pressure was measured using ruby fluorescence. The transition width in zero applied magnetic field, as observed in other high-T_c hydride systems, has a $\Delta T/T_c$ value ranging from 0.005 to 0.036, and such features are readily explained within the dirty limit as described by GL theory[42]. The *V*–*I* relations are simultaneously measured along with the four-probe electrical-resistance measurements. Figure 2b shows the *V*–*I* characteristics with steadily increasing current measured in zero applied magnetic field. At a temperature above the superconducting transition ($T = 297$ K $> T_c = 294$ K), a linear *V*–*I* response following Ohm's law was observed for the metallic state of nitrogen-doped lutetium hydride at 10 kbar. As the temperature is reduced below T_c ($T = 30$ K) at the same pressure, the voltage drop is immeasurably low (essentially zero) and shows a nonlinear, $V \propto I^{3(2.84)}$ response. The *V*–*I* data were obtained from the *V*, *I* pair, shown in Fig. 2b.

Magnetic susceptibility

Another key criterion for a superconducting material is the demonstration of the Meissner effect. In a type I superconductor, this should—in principle—equate to perfect diamagnetism below the critical field, whereas in a type II superconductor, there exists perfect diamagnetism below the lower critical field and a vortex state between the upper

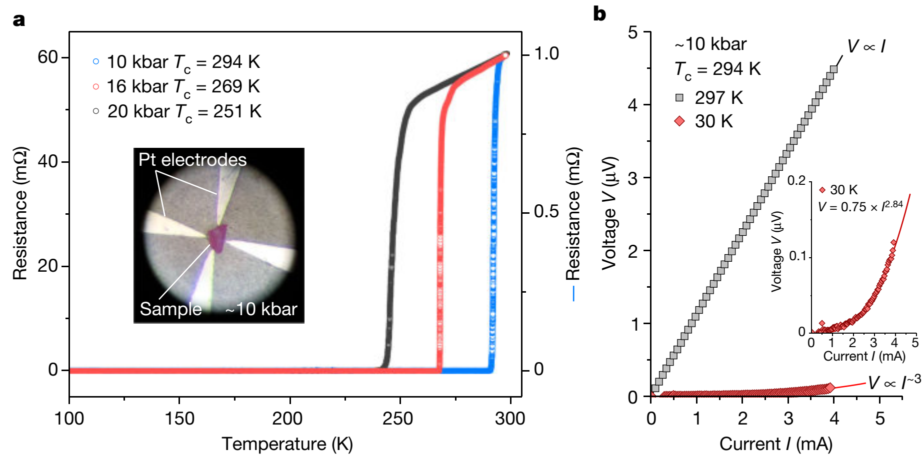


Fig. 2 | Temperature-dependent and field-dependent electrical resistance and *V*–*I* behaviour of the lutetium–nitrogen–hydrogen system. **a**, Temperature-dependent electrical resistance of nitrogen-doped lutetium hydride at high pressures, showing the superconducting transitions as high as 294 K at 10 ± 0.1 kbar, the highest transition temperature measured in all experimental runs. The colours represent the transition at different pressures. The right *y* axis represents the 10-kbar resistance versus temperature in blue colour for a different pair of *V* and *I* contacts. The data were obtained during the warming cycle to minimize the electronic and cooling noise. The inset illustrates the experimental setup for the electrical resistance at around 10 kbar using platinum (Pt) metal probes in a four-probe configuration. **b**, *V*–*I* characteristics measured at temperatures of 297 K and 30 K at 10 kbar. The transition temperature (T_c) is 294 K and above that temperature, the sample exhibits typical linear behaviour. Below T_c (T = 294 K), well within the superconducting regime, at the same pressure, a nonlinear, $V \propto I^{3(2.84)}$ response is shown. The inset highlights the power law *V*–*I* dependence.

and lower critical fields. The temperature dependence of the magnetic moments and *M*–*H* curves at different temperatures were measured on a Quantum Design Physical Property Measurement System (PPMS) by using the Vibrating Sample Magnetometer (VSM) method. Figure 3a shows the d.c. magnetic susceptibility ($\chi = M/H$, in which *M* is magnetization and *H* is magnetic field) as a function of temperature, under conditions of zero field cooling (ZFC) and field cooling (FC) at 60 Oe. The existence of the superconducting phase was then confirmed by measuring the Meissner effect on cooling in a magnetic field. The onset of a well-defined Meissner effect was observed at about 277 K at around 8 kbar. The *M*–*H* data were recorded using a PPMS with VSM option (see Fig. 3b). We used an HMD high-pressure cell and Daphne oil as the medium for applying pressure. Compared with a diamond anvil cell (DAC), the maximum pressure achievable is considerably lower in a HMD high-pressure cell. The HMD high-pressure cell is rated to reach a maximum pressure of 13 kbar, although we have not been able to achieve such a pressure. Because the filling factor of the sample is small (limited by the synthesis procedure), achieving the rated 10 kbar will require substantially greater pressure-cell compression. The maximum pressure we were able to generate was about 8 kbar. The pressure dependence of T_c is approximately 30 K kbar^{-1} from about 3 kbar to about 10 kbar. The broad transition is most probably because of the pressure gradient caused by the high-pressure cell and/or by chemical inhomogeneities in the main sample. The temperatures observed for the onset of diamagnetism with ZFC are commensurate with those from the electrical-resistance measurements.

The diamagnetic response of nitrogen-doped lutetium hydride is also seen by a 30–50-nV drop in the real part of temperature-dependent a.c. susceptibility, $\chi'(T)$, for different pressures across the superconducting stability range of phase II (Fig. 3c). The techniques for measuring a.c. susceptibility were similar to those of Snider et al.[14,15] and with respect to background subtraction; here we have used a cubic polynomial (see Extended Data Fig. 5). Given that the trivalent rare-earth elements are extremely reactive, the synthesis can be tricky and, consequently, the amount of superconducting sample can vary and thus the strength of the signal is dependent on volume. On average, sample sizes are on the order of 70–100 μm in diameter and 10–20 μm thick. For larger samples, the strength of the diamagnetic response increases by nearly five times in magnitude (see Extended Data Fig. 6). The transition width (ΔT_c), as observed in other high-T_c hydride systems, has a value of approximately 0.8 K, 2 K and 5 K at T_c of 294 K (at 10 kbar), 269 K (16 kbar) and 238 K (22 kbar), respectively, indicating the pressure broadening. Extended Data Fig. 12 shows the a.c. susceptibility, with the electrical resistance showing similar critical temperatures and very similar transition widths.

a.c. calorimetry of lutetium–nitrogen–hydrogen

The specific heat (*C*) is an important thermodynamic quantity and is extensively used at ambient pressures to confirm bulk superconductivity. The BCS model superconductors have an energy gap associated with the formation of Cooper pairs, resulting in a spike in the specific heat of a superconductor at T_c. However, detecting such heat anomalies at high pressure is difficult because of the highly thermally conductive environment of the diamonds. In this study, we have used a new setup of the a.c. calorimetric technique for measuring the specific heat[43]. The sample is thermally excited by an a.c. applied at frequency $\omega/2$ to a resistance heater, leading to an a.c. heat power at frequency ω to determine the specific heat capacity (*C*) of the sample (see Methods for more information). Using the well-known superconducting transition in MgB_2 as a test case (Extended Data Fig. 4), we find that the higher-frequency value on the falling edge of the plateau provides the cleanest response for measurements. A frequency sweep above T_c easily identifies the falling edge of the plateau, and the resultant heat capacity of MgB_2 at 15 kbar shows the distinctive discontinuity of the specific heat capacity at 32 K. Replicating this procedure in the superconducting stability regime of phase II yields specific-heat-capacity curves exhibiting the distinctive discontinuity (Fig. 4a–c). The T_c values identified in this manner are very similar to those found through both electrical resistance and a.c. susceptibility, identifying a bulk nature for the transition and the shape of the superconducting dome shown in Fig. 1a. The highest T_c we have observed using this method is 292 K at about 10 kbar. These measurements are also subject to a sample-size dependency, and the typical samples here are about 60–100 μm in diameter.

Composition and structure

The composition of the superconducting ternary lutetium–nitrogen–hydrogen compound was identified by using a combination of

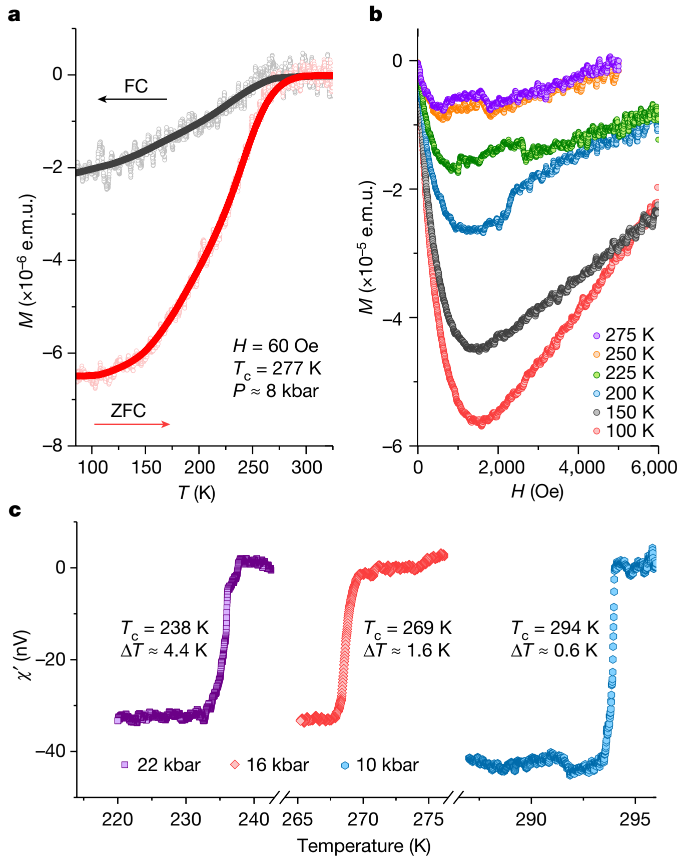


Fig. 3 | Magnetic susceptibility. a, Magnetic susceptibility ($\chi = M/H$, in which M is magnetization and H is magnetic field) as a function of temperature (T) under conditions of zero field cooling (ZFC) and field cooling (FC) at a d.c. field of 60 Oe. **b**, M–H curves recorded close to zero field. **c**, a.c. susceptibility (χ) in nanovolts versus temperature at select pressures, showing marked diamagnetic shielding of the superconducting transition for pressures of 10–22 kbar. The superconducting transition shifts rapidly under pressure to lower temperatures. T_c is determined from the temperature at the onset of the transition. A cubic or quadratic fit of the background signal has been subtracted from the data. We have applied a ten-point adjacent average smoothing for all d.c. magnetization data.

elemental analysis and EDX analysis on the synthesized sample at ambient pressures. The bulk material consistently shows the presence of nitrogen with an average weight percent of 0.8–0.9% N using elemental analysis. EDX, although qualitative, provides evidence of nitrogen in various domains of the overall sample, which we related to partial inhomogeneity in our samples (see Extended Data Fig. 7). The atomic arrangement of these elements is the crux of why the material superconducts, and both XRD and Raman spectroscopy (see Extended Data Fig. 1) show the presence of two distinct hydride compounds in nearly all samples. Both compounds have the same chemical construct of a face-centred cubic (fcc) metal sublattice but with varying contents of hydrogen and nitrogen and a distinct difference in colour. Compound A is indexed as $Fm\bar{3}m$ with a lattice constant of $a = 5.0289(4)$ Å, whereas compound B has $a = 4.7529(9)$ Å, both of which are indicative of lanthanoid hydride and nitride compounds, respectively[9,35,41,44] (Fig. 5a). The mononitride of lutetium is known and reported to adopt a rock-salt (RS) structure with a lattice constant of $a = 4.76$ Å (ref. [44]), very similar to what is measured here for compound B. DFT optimization of the RS and a hypothetical zincblende (ZB) mononitride yields $a = 4.767$ and 5.144 Å, respectively. In the same lattices, the hypothetical RS and ZB structure Lu monohydrides have DFT-optimized lattice constants of $a = 4.800$ and 5.027 Å, respectively. Altering the H and N concentration in the RS structure causes the lattice constant to vary between the limits set by the two parent compounds. Forming fully stoichiometric monohydrides and mononitrides is known to be difficult, so compound B is tentatively assigned as the RS mononitride with potential hydrogen substitution/intercalation, that is, $LuN_{1-\delta}H_{\varepsilon}$.

The electronic and structural evolutions of compound A are codependent and evolve synchronously. Raman spectroscopy of the starting material shows a linear-like progression with compression as modes harden associated with bond stiffening. A gradient change in the mode progression is observed above 3 kbar, in step with the onset of superconductivity in phase II (see Extended Data Fig. 1). There is no registered change in the indexed $Fm\bar{3}m$ symmetry as the phase I to phase II boundary is crossed, indicative of an isostructural second-order phase transition, at least in relation to the cation positions. Compressing phase II into non-superconducting phase III is a first-order structural phase transition with a roughly 0.3% volume discontinuity and a reduction in symmetry of the metal sublattice to the orthorhombic $Immm$ space group. Diminishing quantum properties as the loss of the superconductivity can be associated with a sudden increase in the degrees of freedom associated with phonon propagation through the lattice, yet there is no pronounced change in the Raman spectra other than the expected hardening of modes with compression and the disappearance of one of the low-frequency modes.

At ambient, the dihydride is a blue compound that takes on the fluorite structure (CaF_2, a fcc metal sublattice and the hydrogens almost entirely occupying the tetrahedral interstices), with a lattice constant of 5.033 Å (refs. [45–48]). Although 5.033 Å is larger than what is measured here for compound A, the lanthanoid dihydrides including Lu form solid solutions, LnH_{2+x}, that contract the lattice with further hydrogen uptake[45,46,49,50]. The presence of impurities such as O or N enables LuH_{2+x} to take on extra hydrogen compared with purer metal samples, although that contraction is only 10% that of other Ln hydrides and not enough to explain the lattice constant observed for phase I of compound A (refs. [49,50]). The ambient trihydride adopts a hexagonal $P\bar{3}c1$ phase[51]; however, a cubic trihydride phase begins to form under compression around 12 GPa and is recoverable to ambient[52]. DFT optimizations of the cubic Lu dihydride and trihydride give $a = 5.025$ and 5.012 Å, respectively, and an alternative higher-energy model of the dihydride with 50% of the hydrogens occupying all the octahedral sites and the other 50% occupying the tetrahedral interstices as in the ZB structure gives $a = 4.960$ Å. The disagreement between the computed and experimental lattice constant for the dihydride allows an estimate for the error in the DFT lattice predictions, but—more importantly—the difference in the computed lattice constants corroborates the expected lattice contraction on increasing hydrogen content from the dihydride to the trihydride. We interpret the ZB monohydride structure having a nearly identical lattice constant to the dihydride to mean that little change in the lattice should be anticipated when occupying further tetrahedral interstices with H beyond the ZB structure. For N incorporation into the lattice, as opposed to in grain boundaries as in the studies on further H uptake in the dihydride[49,50], DFT predicts $a = 4.949$ Å for adding a single N at an octahedral interstice in the cubic dihydride lattice, $LuH_2N_{0.25}$, and $a = 5.034$ Å for complete N occupation of the octahedral sites, LuH_2N. In the cubic-trihydride structure, replacing a single H for a N gives $a = 5.028$ and 5.146 Å for octahedral and tetrahedral interstices, respectively, which reduces to $a = 5.021$ and 5.080 Å, respectively, if the single N replacement is in a $2 \times 2 \times 2$ supercell of the primitive rhombohedral cell.

In the harmonic approximation, the dihydride is the only stoichiometric low-H cubic Lu hydride found to be dynamically stable at ambient pressure (see Extended Data Fig. 8). The temperature-dependent conductivity of ambient-pressure LuH_{2+x} is known[49,50,53] and electron–phonon calculations corroborate 0-kbar LuH_2 as non-superconducting with low values of $\lambda = 0.189$ and $\omega_{log} = 363$ K. ZB LuH has a weak instability, with the acoustic phonons at X precluding a calculation of its

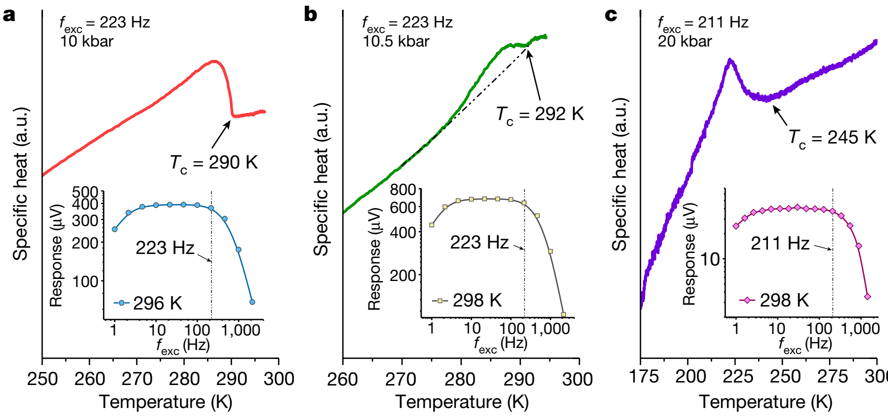


Fig. 4 | Specific-heat-capacity measurement on the superconducting lutetium–nitrogen–hydrogen system. a–c, Specific heat capacity of nitrogen-doped lutetium hydride at 10 kbar (**a**), 10.5 kbar (**b**) and 20 kbar (**c**), showing the superconducting transition as high as 292 K at 10.5 kbar in **b**. The drive frequency (f_{exc}) and frequency sweeps of each measurement are depicted in the insets. The strength of the heat-capacity anomaly associated with superconductivity varied owing to volume fraction as shown in **c**. The dashed line is a guide to the eye to distinguish the trend of the heat-capacity anomaly before and after the transition.

electron–phonon properties without a treatment for anharmonicity, but—because of the similarity between its and LuH_2's phonon band structure—we do not believe it to be a strong candidate for compound A. The larger harmonic dynamic instability at Γ in LuH_3 is a T_{1u} optical phonon mode that splits, with one branch becoming more unstable in other parts of the Brillouin zone. As this lattice could potentially be metastably recovered to ambient conditions[52,54], the stabilization of the material could also be attributed to anharmonic effects. Simulated compression to 15 GPa, at which LuH_3 has been observed experimentally, quenches the harmonic instabilities at the zone centre but not away from it, further indicating that anharmonicity plays a role in stabilizing cubic LuH_3 against the harmonic instabilities of the optical modes. An optical phonon mode of RS LuH also exhibits large harmonic dynamic instabilities, indicating that unstable optical modes arise from hydrogens in the octahedral fcc interstices. This observation is in line with neutron-diffraction results for cubic $NdD_{2.61}$, in which the D in the octahedral site shifts away from the 4b Wyckoff site to a partially occupied, lower-symmetry 32f site[55].

Single N substitution in the cubic cell of LuH_3 at both types of interstitial site increases the (harmonic) dynamic instability at Γ; however, they both split the highly degenerate zone-centred phonon modes of LuH_3, making more of them Raman active, in line with what is observed experimentally. Substitution at the tetrahedral site was found to be more enthalpically favourable than at the octahedral site (not vibrationally corrected), whereas Rietveld refinements provided similar patterns for N substitution at either site (see Extended Data Fig. 9). The LuH_2 to LuH_3 transformation is a metal to semimetal transition, wherein the Lu *d* electron driving the metallicity in LuH_2 donates to/interacts with the octahedral hydrogens in LuH_3, leading to a van Hove singularity[56] just below the Fermi level (see Extended Data Fig. 10). N incorporation at an octahedral interstice sees an increase in metallic character versus LuH_3, with the N p states being metallic, as well as an increased density of H states at the Fermi level. Conversely, N incorporation at a tetrahedral interstice drives the system into a semiconducting state. Although N is more electronegative than H, it does not go to N^{3-} in either case as N p states are in the conduction band, implying *M*–*N* covalencies. Also, N being more electronegative will prevent the formation of only H^- anions in the lattice, and H^- anions are known to be unfavourable for superconductivity. The primary reason for the difference in the electronic behaviours of the two types of substitution is that octahedral substitution has a minimal impact on the parent lattice, whereas tetrahedral substitution pushes the octahedral hydrogens into the opposite octant of the cube, similar to the packing seen in $LaBH_8$ (refs. [21,57,58]) (see Extended Data Fig. 11), which—in turn—leads to disordering of the Lu atoms and an approximately 0.12-Å expansion of the lattice distortions, well beyond what is measured by XRD. A lower N content suppresses the magnitude of the distortions seen for the tetrahedral substitution, and likewise its deviation away from metallicity, but not to the extent to provide a better match with experimental results as compared with octahedral substitution.

Preliminary investigations into adding H-vacancy defects into the cubic cell of the trihydride structure show that, as with the introduction of N, there is a varied response to the lattice, with the removal of an octahedral H giving $a = 5.007$ Å and the removal of a tetrahedral H giving $a = 5.065$ Å. These models both reduce in magnitude but do not remove the optical harmonic dynamic instability at Γ. As with N substitution, these structures split the phonon modes of the parent lattice at Γ, making many of them Raman active, such as what is observed experimentally. Considering that the ground-state structure of several of the ambient rare-earth trihydrides are hexagonal lattices, yet surface defects can metastably trap the high-pressure cubic phase down to ambient conditions as with YH_3 (ref. [54]), the anharmonicity of the hydrogenic phonons[59] along with a combination of N substitutions and H-vacancy defects are probably promoting the formation of a superconducting nitrogen-doped lutetium hydride with higher H content than the dihydride.

Discussion

Clearly, state-of-the-art experiments are needed to determine the exact crystal structure and stoichiometry of nitrogen-doped lutetium hydride and similar materials showing such high-temperature superconducting states. The use of techniques such as neutron diffraction and X-ray spectroscopy, as supported by simulations, are the most likely to provide a route to directly investigating the light elemental content of doped-metal hydrides and to build reliable atomistic descriptions of their chemical environments. A better detailed structural descriptor will enable theoretical modelling of these non-stoichiometric metal hydrides and improved theoretical understanding. An important distinction to make is that XRD was not satisfactorily authoritative for these hydrides even at ambient conditions, a shortcoming of such a technique that we have already highlighted in our work on carbonaceous sulfur hydride[14–16]. The fact that, in this study, we experience a lack of accuracy at ambient conditions confirms our concerns of using XRD techniques for determining hydrogen stoichiometry at more extreme conditions. The inability to accurately measure defect

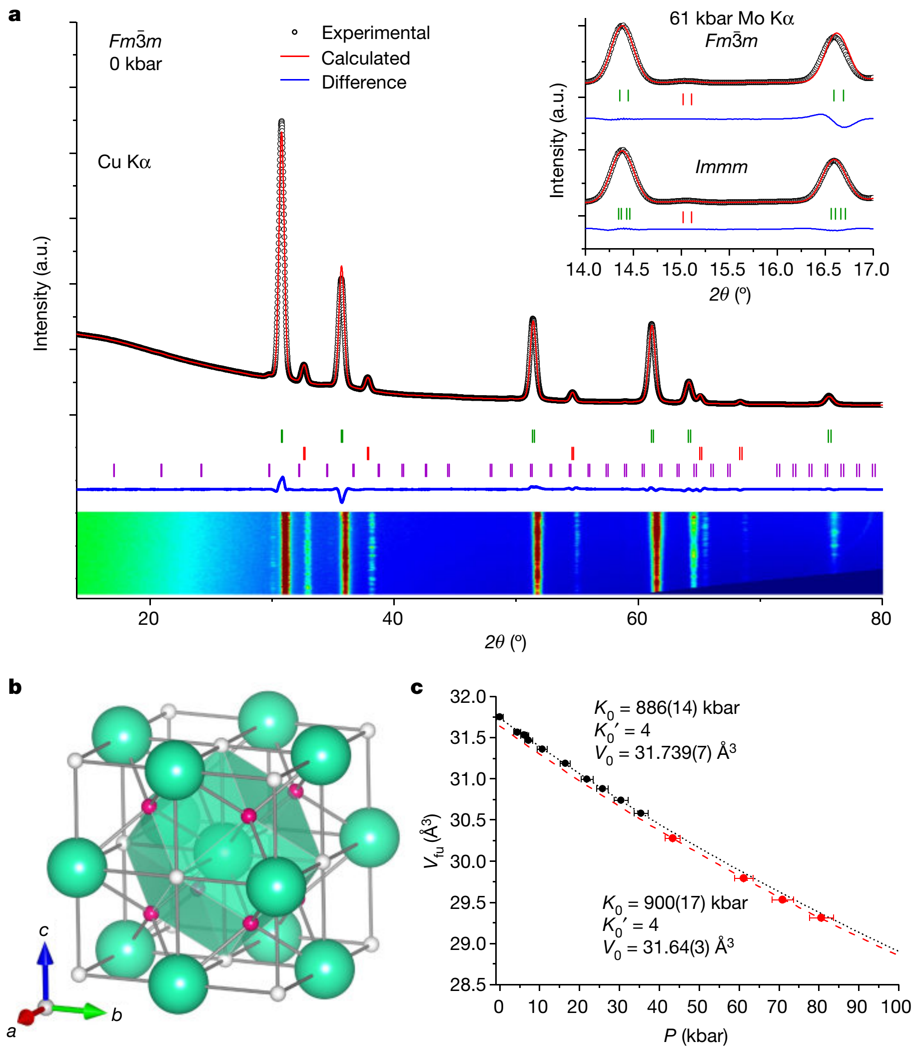


Fig. 5 | XRD studies of the superconducting lutetium–nitrogen–hydrogen system. a, Rietveld refinement of the X-ray powder diffraction data collected at 295 K with Cu Kα radiation. The black points, red line and blue line represent the observed data, calculated intensity and the difference between observed and calculated intensities, respectively. Green tick marks represent the expected Bragg peak positions for the main phase, which is probably $LuH_{3-\delta}N_{\varepsilon}$ (92.25%); minor phases, which are probably $LuN_{1-\delta}H_{\varepsilon}$ (7.29%) and Lu_2O_3 (0.46%), are shown as red and purple tick marks, respectively. The colour map is a cake representation of the X-ray powder diffraction data at ambient pressure. Insets show Le Bail fitting of high-pressure powder diffraction data at 61 kbar with the $Fm\bar{3}m$ and *Immm* space groups. **b**, The crystal structure of the proposed $LuH_{3-\delta}N_{\varepsilon}$ phase. The hydrogens in octahedral interstitial sites are shown in white and those in tetrahedral interstitial sites are in pink. The lutetium atoms are in green and the coordination polyhedron is shown about the central Lu atom. The cell is shifted by (0.5, 0.5, 0.5) fractional coordinates from the standard setting to better represent the coordination polyhedron. **c**, The lattice constant as a function of pressure for the $Fm\bar{3}m$ main phase. The Le Bail method was used for the refinement of the high-pressure XRD data. The equation of state was fitted using the Birch–Murnaghan method, as shown by the dashed lines for two pressure ranges, 0 kbar < P < 40 kbar (black) and P > 42.7 kbar (red), and the corresponding K_0 (bulk modulus at $P = 0$), V_0 (reference volume at $P = 0$) and K_0' (derivative of the bulk modulus with respect to pressure at $P = 0$) are shown.

densities and fractional occupancies of the lightest elements is mostly a consequence of extremely complex synthesis techniques and which in turn limits the accurate boundary conditions to aid theoretical methods for modelling and predicting the quantum properties of such materials. In summary, the most remarkable result of this study is the evidence for the near-ambient superconducting state observed in N-doped lutetium hydride with T_c of 294 K at 10 kbar. On the basis of the measured XRD and Raman spectra, the observed superconducting properties can most probably be attributed to $Fm\bar{3}m$ $LuH_{3-\delta}N_{\varepsilon}$, for which different non-stoichiometric values are used to indicate the possibility of both N-substitution and H-vacancy defects. The physical properties of the superconducting N-doped lutetium hydride will be better constrained by magnetic-field-dependence resistance, susceptibility and heat-capacity measurements. Whilst all other high-temperature superconducting metal hydrides have been observed at multi-megabar pressure conditions, our discovery of a 21 °C superconducting material at 10 kbar will certainly lead to the emergence of a new field of materials science, as such conditions are substantially more accessible to a multitude of new researchers outside the field of high-pressure physics.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-05742-0.

1. Onnes, H. K. The resistance of pure mercury at helium temperatures. *Commun. Phys. Lab. Univ. Leiden* **12**, 120 (1911).
2. Ginzburg, V. L. Nobel Lecture: On superconductivity and superfluidity (what I have and have not managed to do) as well as on the "physical minimum" at the beginning of the XXI century. *Rev. Mod. Phys.* **76**, 981–998 (2004).
3. Bednorz, J. G. & Müller, K. A. Possible high T_c superconductivity in the Ba–La–Cu–O system. *Z. Phys. B Condens. Matter* **64**, 189–193 (1986).
4. Wu, M. K. et al. Superconductivity at 93 K in a new mixed-phase Y-Ba-Cu-O compound system at ambient pressure. *Phys. Rev. Lett.* **58**, 908–910 (1987).
5. Schilling, A., Cantoni, M., Guo, J. D. & Ott, H. R. Superconductivity above 130 K in the Hg–Ba–Ca–Cu–O system. *Nature* **363**, 56–58 (1993).
6. Carlsson, A. E. & Ashcroft, N. W. Approaches for reducing the insulator-metal transition pressure in hydrogen. *Phys. Rev. Lett.* **50**, 1305–1308 (1983).
7. Ashcroft, N. W. Hydrogen dominant metallic alloys: high temperature superconductors? *Phys. Rev. Lett.* **92**, 187002 (2004).
8. Liu, H., Naumov, I. I., Hoffmann, R., Ashcroft, N. W. & Hemley, R. J. Potential high-T_c superconducting lanthanum and yttrium hydrides at high pressure. *Proc. Natl Acad. Sci.* **114**, 6990–6995 (2017).
9. Peng, F. et al. Hydrogen clathrate structures in rare earth hydrides at high pressures: possible route to room-temperature superconductivity. *Phys. Rev. Lett.* **119**, 107001 (2017).
10. Drozdov, A. P. et al. Superconductivity at 250 K in lanthanum hydride under high pressures. *Nature* **569**, 528–531 (2019).
11. Somayazulu, M. et al. Evidence for superconductivity above 260 K in lanthanum superhydride at megabar pressures. *Phys. Rev. Lett.* **122**, 27001 (2019).
12. Snider, E. et al. Synthesis of yttrium superhydride superconductor with a transition temperature up to 262 K by catalytic hydrogenation at high pressures. *Phys. Rev. Lett.* **126**, 117003 (2021).
13. Troyan, I. A. et al. Anomalous high-temperature superconductivity in YH_6. *Adv. Mater.* **33**, 2006832 (2021).
14. Snider, E. et al. Retraction article: Room-temperature superconductivity in a carbonaceous sulfur hydride. *Nature* **586**, 373–377 (2020).
15. Snider, E. et al. Retraction note: Room-temperature superconductivity in a carbonaceous sulfur hydride. *Nature* **610**, 804 (2022).
16. Smith, G. A. et al. Carbon content drives high temperature superconductivity in a carbonaceous sulfur hydride below 100 GPa. *Chem. Commun.* **58**, 9064–9067 (2022).
17. Sun, Y., Lv, J., Xie, Y., Liu, H. & Ma, Y. Route to a superconducting phase above room temperature in electron-doped hydride compounds under high pressure. *Phys. Rev. Lett.* **123**, 097001 (2019).
18. Ge, Y., Zhang, F. & Hemley, R. J. Room-temperature superconductivity in boron- and nitrogen-doped lanthanum superhydride. *Phys. Rev. B* **104**, 214505 (2021).
19. Grockowiak, A. D. et al. Hot hydride superconductivity above 550 K. *Front. Electron. Mater.* **2**, 837651 (2022).
20. Zhang, Z. et al. Design principles for high-temperature superconductors with a hydrogen-based alloy backbone at moderate pressure. *Phys. Rev. Lett.* **128**, 047001 (2022).
21. Di Cataldo, S., Heil, C., von der Linden, W. & Boeri, L. $LaBH_8$: towards high-T_c low-pressure superconductivity in ternary superhydrides. *Phys. Rev. B* **104**, L020511 (2021).
22. Ashcroft, N. W. Metallic hydrogen: a high-temperature superconductor? *Phys. Rev. Lett.* **21**, 1748 (1968).
23. Richardson, C. F. & Ashcroft, N. W. High temperature superconductivity in metallic hydrogen: electron-electron enhancements. *Phys. Rev. Lett.* **78**, 118–121 (1997).
24. Dias, R. P. & Silvera, I. F. Observation of the Wigner-Huntington transition to metallic hydrogen. *Science* **355**, 715–718 (2017).
25. Loubeyre, P., Occelli, F. & Dumas, P. Synchrotron infrared spectroscopic evidence of the probable transition to metal hydrogen. *Nature* **577**, 631–635 (2020).
26. Wang, H., Tse, J. S., Tanaka, K., Iitaka, T. & Ma, Y. Superconductive sodalite-like clathrate calcium hydride at high pressures. *Proc. Natl Acad. Sci.* **109**, 6463–6466 (2012).
27. Drozdov, A. P., Eremets, M. I., Troyan, I. A., Ksenofontov, V. & Shylin, S. I. Conventional superconductivity at 203 kelvin at high pressures in the sulfur hydride system. *Nature* **525**, 73–76 (2015).
28. Bi, T., Zarifi, N., Terpstra, T. & Zurek, E. in *Reference Module in Chemistry, Molecular Sciences and Chemical Engineering* (Elsevier, 2019).
29. Hilleke, K. P. & Zurek, E. Tuning chemical precompression: theoretical design and crystal chemistry of novel hydrides in the quest for warm and light superconductivity at ambient pressures. *J. Appl. Phys.* **131**, 070901 (2022).
30. Di Cataldo, S., von der Linden, W. & Boeri, L. First-principles search of hot superconductivity in La-*X*-H ternary hydrides. *npj Comput. Mater.* **8**, 2 (2022).
31. Di Cataldo, S., Qulaghasi, S., Bachelet, G. B. & Boeri, L. High-T_c superconductivity in doped boron-carbon clathrates. *Phys. Rev. B* **105**, 064516 (2022).
32. Ye, X., Zarifi, N., Zurek, E., Hoffmann, R. & Ashcroft, N. W. High hydrides of scandium under pressure: potential superconductors. *J. Phys. Chem. C* **122**, 6298–6309 (2018).
33. Shannon, R. D. Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Crystallogr. A* **32**, 751–767 (1976).
34. Rumble, J. R. (ed.) *CRC Handbook of Chemistry and Physics* 102nd edn (CRC Press/Taylor & Francis, 2021).
35. Greenwood, N. N. & Earnshaw, A. (eds) *Chemistry of the Elements* 2nd edn (Butterworth-Heinemann, 1997).
36. Zhou, D. et al. Superconducting praseodymium superhydrides. *Sci. Adv.* **6**, 6849–6877 (2020).
37. Zhou, D. et al. High-pressure synthesis of magnetic neodymium polyhydrides. *J. Am. Chem. Soc.* **142**, 2803–2811 (2020).
38. Semenok, D. V. et al. Effect of magnetic impurities on superconductivity in LaH_{10}. *Adv. Mater.* **34**, 2204038 (2022).
39. Sun, W., Kuang, X., Keen, H. D. J., Lu, C. & Hermann, A. Second group of high-pressure high-temperature lanthanide polyhydride superconductors. *Phys. Rev. B* **102**, 144524 (2020).
40. Jaroń, T. et al. Synthesis, structure, and electric conductivity of higher hydrides of ytterbium at high pressure. *Inorg. Chem.* **61**, 8694–8702 (2022).
41. Song, H. et al. High T_c superconductivity in heavy rare earth hydrides. *Chin. Phys. Lett.* **38**, 107401 (2021).
42. Cornelius, A. L., Lawler, K. V. & Salamat, A. Understanding hydrogen rich superconductors: importance of effective mass and dirty limit. Preprint at https://doi.org/10.48550/arxiv.2202.04254 (2022).
43. Dasenbrock-Gammon, N., McBride, R., Yoo, G., Dissanayake, S. & Dias, R. Second harmonic AC calorimetry technique within a diamond anvil cell. *Rev. Sci. Instrum.* **93**, 093901 (2022).
44. Klesnar, H. P. & Rogl, P. Phase relations in the ternary systems rare-earth metal (RE)-boron-nitrogen, where RE = Tb, Dy, Ho, Er, Tm, Lu, Sc and Y. *High Temp. High Press.* **22**, 453–457 (1990).
45. Pebler, A. & Wallace, W. E. Crystal structures of some lanthanide hydrides. *J. Phys. Chem.* **66**, 148–151 (1962).
46. Bonnet, J. E. & Daou, J. N. Rare-earth dihydride compounds: lattice thermal expansion and investigation of the thermal dissociation. *J. Appl. Phys.* **48**, 964–968 (1977).
47. Weaver, J. H., Rosei, R. & Peterson, D. T. Electronic structure of metal hydrides. I. Optical studies of ScH_2, YH_2, and LuH_2. *Phys. Rev. B* **19**, 4855–4866 (1979).
48. Peterman, D. J., Harmon, B. N., Marchiando, J. & Weaver, J. H. Electronic structure of metal hydrides. II. Band theory of ScH_2 and YH_2. *Phys. Rev. B* **19**, 4867–4875 (1979).
49. Daou, J. N., Vajda, P., Burger, J. P. & Shaltiel, D. Percolating electrical conductivity in two phased LuH_{2+x} compounds. *Europhys. Lett.* **6**, 647–651 (1988).
50. Daou, J. N., Vajda, P., Burger, J. P. & Shaltiel, D. Percolating electrical conductivity in two phased LuH2+x compounds. *Europhys. Lett.* **8**, 587 (1989).
51. Mansmann, M. & Wallace, W. E. The structure of HoD3. *J. Phys.* **25**, 454–459 (1964).
52. Palasyuk, T. & Tkacz, M. Pressure-induced structural phase transition in rare-earth trihydrides. Part I. (GdH_3, HoH_3, LuH_3). *Solid State Commun.* **133**, 481–486 (2005).
53. Daou, J. N., Lucasson, A., Vajda, P. & Burger, J. P. Observation of the optical and acoustic electron-phonon coupling in Sc, Y and Lu dihydrides and dideuterides by electrical resistivity. *J. Phys. F Metal Phys.* **14**, 2983–2993 (1984).
54. Kataoka, R. et al. The origin of the highly crystallized face-centered cubic YH_3 high-pressure phase when quenched to ambient condition. *Mater. Today Commun.* **31**, 103265 (2022).
55. Renaudin, G., Fischer, P. & Yvon, K. Neodymium trihydride, NdH_3, with tysonite type structure. *J. Alloys Compd.* **313**, L10–L14 (2000).
56. Villa-Cortés, S. & De la Peña-Seaman, O. Effect of van Hove singularity on the isotope effect and critical temperature of H_3S hydride superconductor as a function of pressure. *J. Phys. Chem. Solids* **161**, 110451 (2022).
57. Liang, X. et al. Prediction of high-T_c superconductivity in ternary lanthanum borohydrides. *Phys. Rev. B* **104**, 134501 (2021).
58. Belli, F. & Errea, I. Impact of ionic quantum fluctuations on the thermodynamic stability and superconductivity of. *Phys. Rev. B* **106**, 134509 (2022).
59. Errea, I. Superconducting hydrides on a quantum landscape. *J. Phys. Condens. Matter* **34**, 231501 (2022).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2023

Methods

High-pressure experiments

This study was based on a large number of experiments, with more than a hundred samples. The samples were loaded onto a membrane-driven diamond anvil cell (m-DAC), using 1/3-carat, type Ia diamond anvils with a 0.2-mm, 0.4-mm, 0.6-mm and 0.8-mm (for low pressures) culet. A 0.25-mm-thick rhenium gasket was pre-indented to 15–100 μm (depending on the pressure and experiment) and a 120-μm hole (or 280 μm or 600 μm for low pressures) was electrospark drilled at the centre of the gasket. We used a high-pressure gas loader to compress gasses to high densities. The m-DACs were first loosely closed and mounted into a gearbox. The DAC and gearbox were then placed into a high-pressure gas loader (Top Industries). The system was first flushed with gasses to purge the circuit of impurities and then the sample chamber was pressurized. A 100-μm Lu foil was compressed between two diamonds to make it thinner than 100 μm and then loaded into a DAC with the H_2/N_2 gas mixture (99:1) and pressed to 2 GPa. We have not used any other gas ratios in our synthesis. All of the prepared samples were kept in a glovebox, closed and reopened to load with gas in a pre-purged hydrogen-rich environment. The glovebox environment was operated at levels at which O_2 and H_2O were each less than 0.5 ppm. The sample was heated in an oven overnight at 65 °C. After 24 h, the DAC was released to recover the sample. The samples were characterized using Raman and XRD studies. We have also used commercially available LuH_2/LuH_3. Owing to extremely complex synthesis techniques, controlling the correct stoichiometry with the right amount of hydrogen and nitrogen percentages was extremely challenging. The success rate of measuring a sample with superconducting properties was about 35%.

Raman spectroscopy, pressure and temperature determination

Raman spectroscopy was carried out using a custom micro-Raman setup in back-scattering geometry, along with Bragg notch filters, using a 532-nm Millennia eV laser, a Princeton Instruments HRS-500 spectrometer and a Pylon camera. Pressure determination was predominantly carried out using ruby fluorescence with the pressure gauge of Shen et al.[60] and the temperature correction of Datchi et al.[61]. Furthermore, pressure was also measured from the observed Raman peak location following several calibrations runs to determine the pressure dependence of the Raman modes. Pressure was measured during both cooling and heating. DT-670 silicon diodes were used to measure the temperature to five digits with our temperature controller (that is, two decimal places above 100 K or three decimal places below 100 K). The temperature uncertainty is larger than that precision as the diodes cannot be placed directly on the sample in a DAC and are placed outside the sample chamber and on the mechanical assembly around the DAC. Thus, the temperature uncertainty is dominated by thermal gradients between the temperature probes and the sample in the DAC, which are large while cooling down but much smaller while warming owing to the associated cooling and warming rates.

Electrical-resistance measurements

The resistance measurements were performed using standard DAC techniques with a standard four-probe technique similar to Dias et al.[62] that was used to measure the resistance of the sample. Either Al_2O_3 or diamond powder was packed into an insulating shell in which sample is loaded. Platinum foil (5 μm in thickness) cut electrodes are placed in contact with the sample, leading out of the pressure cell, allowing for transport measurements of the samples under pressure. Electrical resistance is thus measured in a four-probe configuration. The resistance measured on both warming and cooling at about 10 kbar is shown in Extended Data Fig. 13. A Keithley 6221 current source is used to apply a low-frequency (13 or 17 Hz) current across two of the probes and a SR860 lock-in amplifier with a 500× preamplifier measures the resulting voltage across the remaining two probes. For relatively high (on the order of about kΩ or more) resistance samples, a d.c. configuration is used, whereby a d.c. current is applied across two of the probes using a SRS CS580 current source, whereas the resulting voltage on the remaining two probes is measured with a DMM6500. In some cases, small residual resistance from the instrument offsets was subtracted from the measured voltage. We used a custom-built BeCu DAC for magnetic-field-dependent electrical-resistance studies using a custom-designed magnet from ColdEdge Inc. The *V*–*I* curves were obtained by supplying an a.c. current between 0 and 4 mA across the sample and measuring the voltage response, at temperatures above and below the critical temperature T_c. Above T_c, a current sweep from 0 to 4 mA was performed in 50 equally spaced steps and a linear response in the *V*–*I* curve is observed.

a.c. calorimetry measurements

For measurement of the specific heat capacity, a modified version of the conductivity setup is used[43]. Two sets of two probes are shorted in contact with the sample. The first, Ti, nichrome or Pt, serves as the heater. The second uses a pair of materials with differing Seebeck coefficients, either Pt/Ag or the standard alumel/chromel thermocouple pair. Typically, a NaCl insulating insert was used to isolate the sample better thermally. Extended Data Fig. 2, top shows the schematic representation of the setup and Extended Data Fig. 2, bottom shows the actual setup in the DAC.

To make measurements, the heater is driven at frequency *f* (typically between 50 and 500 Hz). The sample is thus ohmically heated and modulated at frequency $2 \times f$. Then the DAC is externally cooled or heated and the voltage measured across the thermocouple, using a lock-in amplifier reading at frequency $2 \times f$, which was shown to inversely proportional to the specific heat of the sample at any given temperature[63]. To verify appropriate behaviour, both current and frequency sweeps of the drive are performed, as shown in Extended Data Fig. 3. Frequency dependence is shown to exhibit a characteristic dome shape with a wide plateau, and the current sweep shows a quadratic dependence, as expected from ohmic heating. For this type of a.c. calorimetric measurements, the governing equation is given by[63]: $C_{\mathrm{a.c}} = \frac{P}{\omega T_{\mathrm{a.c.}}} F(\omega)$, with $C_{\mathrm{a.c.}}$ being the a.c. heat capacity, P the driving power, ω the driving frequency, $T_{\mathrm{a.c.}}$ the modulation in temperature of the sample and $F(\omega)$ the frequency response curve. The details of the frequency response curve are described elsewhere[63,64] but it depends on the relation between the three timescales of the system: $\tau_1 = 1/\omega_1$ is the timescale over which the sample thermalizes with the environment, that is, how fast heat dissipates from the sample to the NaCl in which the sample is sitting; $\tau_2 = 1/\omega_2$ is the timescale over which the heater, sample and thermocouple thermalize with each other; and $\tau_3 = 1/\omega_3$ is the timescale of the driving heater. The first two timescales, ω_1 and ω_2, are determined by the exact details of the experimental setup and will vary between runs, although care is taken so that $\omega_1 < \omega_2$ and the sample is able to thermalize with the heater and thermocouple before heat dissipates into the environment. This is accomplished by using the thermally insulating NaCl medium.

The third timescale is chosen by the frequency of the driving current and is an experimental parameter. Before performing an experiment, care is taken so that an appropriate driving frequency, or ω_3, is chosen. This is best seen by performing a frequency sweep of the drive, as shown in Extended Data Fig. 3, and shows the frequency response curve, $F(\omega)$. The response falls into three regions: an initial rise, a flat plateau and a final decline. The first region corresponds to $\omega_3 \ll \omega_1$; low-frequency drive results in heat dissipating to the environment, reducing the change in temperature of the sample. The third region corresponds to $\omega_3 \gg \omega_2$; the sample is not able to thermalize with the heater/thermocouple because the drive is too fast. The intermediate plateau between these two regions is where the response function is relatively constant with frequency, and there is good coupling between the heater, sample and thermocouple, with minimized loss to the environment. The final consideration is experimental. By using higher

frequencies, the signal-to-noise ratio is improved and so a frequency between regions two and three is typically chosen.

Because the heater is driven at frequency f, the sample will experience temperature oscillations at frequency $2 \times f$ and the thermocouple will produce a voltage at $2 \times f$. The heat capacity is then inversely proportional to the measured voltage. As extra verification, a current sweep is performed to make sure the expected quadratic behaviour is produced, as shown in Extended Data Fig. 3, inset. A current is chosen at which acceptable signal is measured, but high currents are avoided to minimize d.c. heating that may occur on the sample. An example measurement and frequency sweep are shown in Extended Data Fig. 4.

a.c. magnetic-susceptibility measurements

All experiments were performed using a side-by-side double-coil technique as described by Debessai et al.[65]. In a single-coil setup, the measured voltage will be proportional to the average magnetic susceptibility of the volume contained in the coil. Owing to the geometry of the diamonds used within a DAC, the sample is necessarily a small fraction of the volume contained in the coil. We therefore use a double-coil setup whereby a second coil is connected in series but reversed relative to the first coil. We call the coil surrounding the sample the 'primary' coil and the second coil the 'dummy' coil. When connected in reverse, the signal from both coils essentially subtracts. Because the coils are near identical, when subtracted, the remaining difference should be because of the sample. In practice, the coils are never perfectly identical. The coils are hand wound in house to be as identical as possible, before being balanced for use in an experiment. To balance a pair of coils, first the pickup from the primary coil is measured and then the second coil is connected in reverse. While monitoring the pickup, one of the coils is slowly unwound until the measured pickup is as small as possible. The balancing is done until the signal is less than 1% of the single-coil reference value and most coils are balanced to 0.1–0.5%. This residual signal is what is measured as the (relatively) large background as compared with the signal strength coming from the sample. For improved resolution of the small signal voltages, a 500× preamplifier (SR554 Transformer Preamplifier) is often used. After dividing out the 500× preamplifier, the signal strengths are in the range of approximately 10–200 nV, depending on the sample and conditions. Owing to the small sample size, a large temperature-dependent background signal is observed, but the transition is clearly visible despite the large magnitude of the background. A background must be subtracted from the real part, yielding the final susceptibility signal. In this work, cubic or quadratic polynomial backgrounds were used (Extended Data Fig. 5), taking the measured voltage either immediately before or immediately after the transition.

d.c. magnetic susceptibility

We used the non-membrane-driven clamp-style pressure cell manufactured by HMD, a leading Japanese supplier of pressure cells for magnetometry. A schematic and actual picture of the HMD-13 cell can be found in the cell manual. This simplified design requires neither copper sealing rings nor a hydraulic press to achieve the pressure. The HMD-13 cell is designed with a BeCu construction and affords a minimized, uniform magnetic background, typically 4×10^{-7} e.m.u. gauss^{-1}. This pressure cell is a clamp cell, in which the sample is loaded with Daphne 7373 pressure medium inside a Teflon tube and closed using a Teflon capsule. Measurements were performed by directly measuring the HMD cell, which does not require a diamond or gasket. The samples that we use for this measurement are much larger compared with the DAC measurements. We have loaded large (approximately 150 μm × 100 μm) pieces into the Teflon capsule. The sample centre is known from the dimensions and pressure-cell compression, which is very standard in these experiments. The magnetization measurement was performed using a VSM option. The VSM option includes a linear motor transport for vibrating the sample, a coil set and the software application. An empty cell without the sample while keeping everything else the same was also performed to identify the cell background. The empty cell background is mostly constant with temperature. A linear or cubic background response was subtracted from the data, and we have applied a ten-point adjacent-average smoothing for all of the data (see Extended Data Fig. 14). The HMD-13 high-pressure cell is rated to reach a maximum pressure of up to 13 kbar. To achieve the highest pressure rated for this cell, the Teflon sample tube should be filled with as much sample as possible, with the minimum amount of pressure-transmitting medium required to fill the Teflon sample chamber. If the filling factor of the sample is small (limited by the synthesis procedure), achieving the rated pressure of 10 kbar will require substantially greater pressure-cell compression.

XRD and elemental analysis

For ambient X-ray powder diffraction measurements, microgram polycrystalline samples with typical sizes ($0.07 \times 0.05 \times 0.02$ mm^3) were placed onto a nylon loop and mounted in a Rigaku XtaLAB Synergy-S Dualflex diffractometer equipped with a HyPix-6000HE Hybrid X-ray Photon Counting area detector. The full data collection was carried out using a PhotonJet (Cu Kα) X-ray source with a detector distance of 34.0 mm. Data collection was performed using Gandolfi scans, which randomize the sample orientation in the beam by driving both phi and omega circles simultaneously. The scans were repeated for different kappa settings. CrysAlis Pro software was used to process and evaluate powder measurements. The Rietveld refinements of the ambient XRD data were performed using FullProf. High-pressure XRD measurements were performed using the Rigaku high-pressure kit designed for the Rigaku XtaLAB Synergy-S Dualflex diffractometer. Pressure was generated in custom-made PEAS-Q36 DACs. Two-hundred-micrometre conical diamonds were mounted on tungsten carbide bases with 70° opening angle. Data collection was carried out using a PhotonJet (Mo Kα) X-ray source with a detector distance of 80 mm. Samples were placed into either a rhenium or a tungsten gasket with a glycerine or methanol/ethanol pressure medium. Ag pieces (roughly 20 × 20 μm) were placed with the samples as a pressure marker. Pressure was estimated using the equation of state of Ag at 295 K. The pressure dependence of the lattice parameters was obtained from Le Bail refinement for the high-pressure XRD data using FullProf.

Elemental analysis was carried out using a PerkinElmer 2400 Series II CHNS/O Elemental Analyzer instrument for rapid determination of the carbon, hydrogen and nitrogen content in our samples. The instrument uses a helium carrier gas with an accuracy of about 0.3% for each element. The samples are crimp-sealed in special tin capsules before being loaded and burned in the instrument. Several different samples were prepared in an Ar glovebox to compare the elemental analysis results with samples that were prepared in air. Similar N content was detected in the samples tested in Ar atmosphere and air.

Simulations

Plane-wave DFT simulations using the Perdew–Burke–Ernzerhof[66] generalized gradient approximation functional were performed with Quantum ESPRESSO[67,68]. The convergence threshold for self-consistent field energies was 10^{-13} Ry, the convergence for forces was 10^{-6} Ry Bohr^{-1} and the stress convergence was 10^{-3} kbar. Gaussian smearing was used with a smearing width of 0.015 Ry. The phonons of cubic unit cells were calculated on $4 \times 4 \times 4$ $\vec{q}$-grids and those of rhombohedral primitive cells were calculated on $6 \times 6 \times 6$ $\vec{q}$-grids. Rhombohedral primitive cells were used whenever possible. The $\vec{k}$-grid density for the structural optimizations and phonon simulations was double that of the $\vec{q}$-grid in each direction, and the denser $\vec{k}$-grid for electron–phonon couplings was four times as dense as the $\vec{q}$-grid in each direction. In all visualizations and electron–phonon calculations, the 'simple' acoustic sum rule correction implemented in Quantum ESPRESSO was applied to the computed phonons. The PseudoDojo norm-conserving pseudopotentials were used with a kinetic energy cutoff of 100 Ry and charge density cutoff of 400 Ry (ref. [69]). Norm-conserving pseudopotentials were used

to perform the electron–phonon calculations, along with a simplified Hubbard correction (DFT+U) applied to the *f* electrons[70,71]. A Hubbard U of 5.5 eV was selected in a similar fashion to the pseudopotential formulation of Topsakal and Wentzcovitch[72], that is, comparing the optimized lattice constants of the metal mononitride to experiment on a 0.5-eV interval. Those pseudopotentials were not used because they are at present incompatible with Quantum ESPRESSO density functional perturbation theory phonon calculations with a +U correction. The importance of describing the *f* electrons with DFT+U was tested by performing structural optimizations with the *f* electrons in core (using the pslibrary[73] PAW[74] pseudopotential for the metal) and with no Hubbard correction applied. In both of those cases, we found the lattice constants of all evaluated compounds to be underestimated compared with the available literature and experimental results. Neither of those two approaches were found to eliminate the dynamic instabilities of the trihydride (Extended Data Fig. 8b), although the instability at Γ was reduced in magnitude with the *f* electrons placed in core. To explore the dynamic instability of the trihydride, the displacements of NdD_3 (ref. [55]) were tested, as well as adding noise to each H position in the cubic unit cell in line with the displacements along one of the unstable optical phonon modes at Γ. The resulting structure was lower in energy and determined to be of the *Pmnm* space group, a subgroup of the XRD-determined *Immm* phase III structure (Extended Data Fig. 11b). Substitution of a N into a tetrahedral site of the distorted cubic representation of the *Pmnm* structure strongly reduced the distortions away from cubic (that is, cell edges that differ by less than 0.001 Å rather than 0.640 Å). The lighter lanthanoid hydrides $LnH(D)_x$ are known to tetragonally distort above $x \geq 2.3$ (refs. [75,76]), so it is possible that the phase II to III transformation is driven by the harmonic dynamic instabilities of the cubic trihydride. As there is no other hydrogen present and the transformation into phase III is recoverable to phase I, it is not likely that the sample is undergoing disproportionation to form a variant of the predicted tetragonal LuH_4 lattice[41].

It should be noted that LuH_2 has a particularly low uncorrected acoustic phonon frequency of about −110 cm^{-1} at Γ, which is not improved by doubling the size of the $\vec{k}$-grid in each direction or increasing the wavefunction cutoff. We found that, instead of representing the primitive cell of LuH_2 with the more highly symmetric lattice vectors of a fcc system, using a triclinic representation with *x* along ***a*** and *z* along *c** (which uses a D_3 electronic point group as opposed to O_h) has a negligible effect on the optimized lattice but alleviates the very negative uncorrected frequency at Γ. However, this creates weak dynamic instabilities of the acoustic phonons just off of Γ (see Extended Data Fig. 8), implying that anharmonic contributions may play a role in stabilizing the lattice, similar to how they were found to stabilize $Im\bar{3}m$ H_3S below 175 GPa (ref. [77]). Changing between the triclinic or more symmetric representation of the primitive unit cell's lattice vectors does not alleviate the optical dynamic instabilities of LuH_3. In addition, the phonon band dispersions for LuH_2 and ZB LuH in the highly symmetric lattice vectors were evaluated as a function of pressure up to 50 kbar. However, no notable change to their phonon band structures was observed, including the instability at *X* for ZB LuH (Extended Data Fig. 16).

Data availability

The authors declare that the data supporting the findings of this study are available in the article and its supplementary information files and from the public link https://doi.org/10.5281/zenodo.7374510. Source data are provided with this paper.

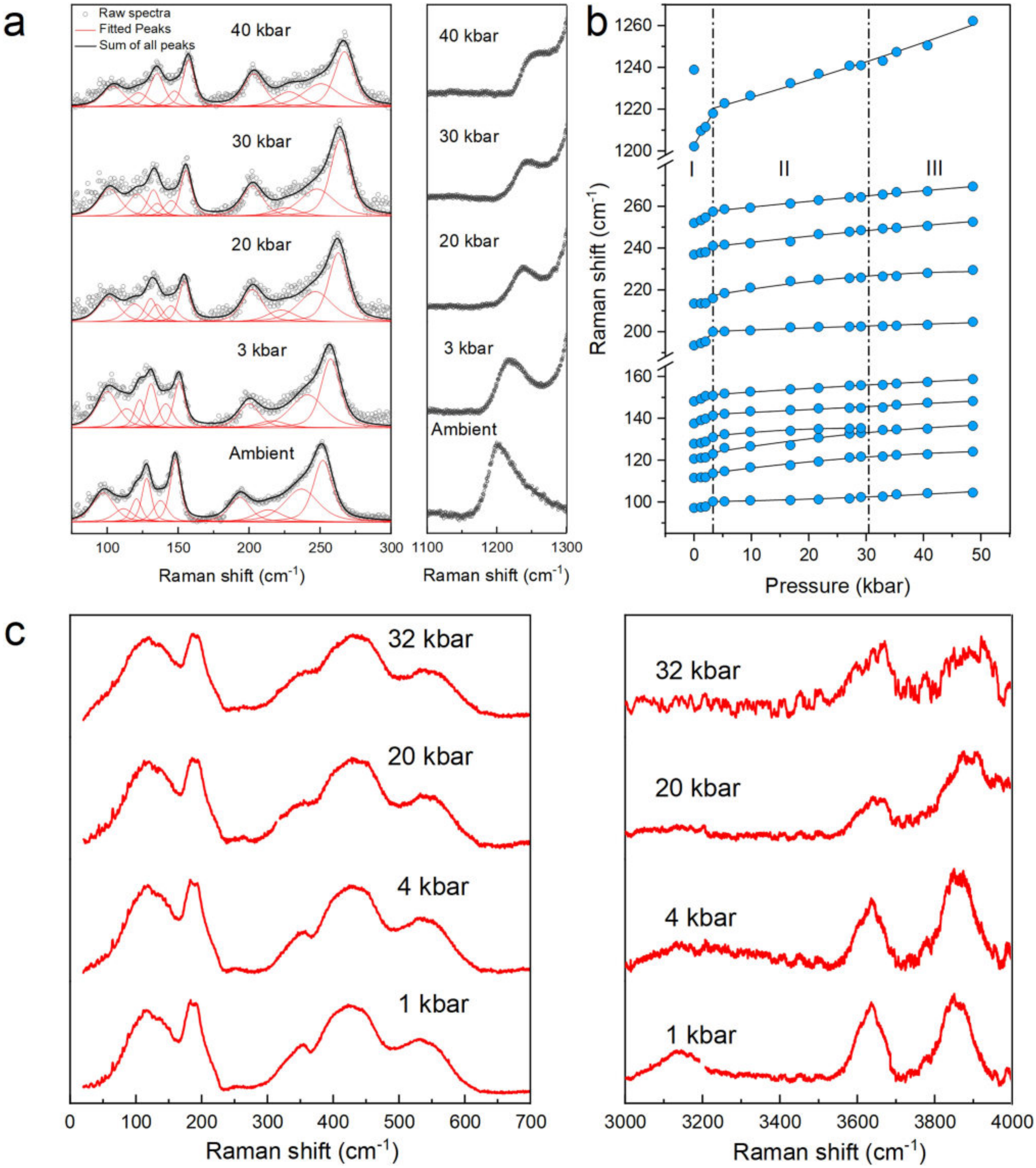
60. Shen, G. et al. Toward an international practical pressure scale: a proposal for an IPPS ruby gauge (IPPS-Ruby2020). *High Press. Res.* **40**, 299–314 (2020).
61. Datchi, F. et al. Optical pressure sensors for high-pressure–high-temperature studies in a diamond anvil cell. *High Press. Res.* **27**, 447–463 (2007).
62. Dias, R. P., Yoo, C.-S., Kim, M. & Tse, J. S. Insulator-metal transition of highly compressed carbon disulfide. *Phys. Rev. B* **84**, 144104 (2011).
63. Li, Y.-S., Borth, R., Hicks, C. W., Mackenzie, A. P. & Nicklas, M. Heat-capacity measurements under uniaxial pressure using a piezo-driven device. *Rev. Sci. Instrum.* **91**, 103903 (2020).
64. Kraftmakher, Y. *Modulation Calorimetry. Theory and Applications* (Springer, 2004).
65. Debessai, M., Hamlin, J. J. & Schilling, J. S. Comparison of the pressure dependences of T_c in the trivalent *d*-electron superconductors Sc, Y, La, and Lu up to megabar pressures. *Phys. Rev. B* **78**, 064519 (2008).
66. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
67. Giannozzi, P. et al. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys. Condens. Matter* **21**, 395502 (2009).
68. Giannozzi, P. et al. Advanced capabilities for materials modelling with Quantum ESPRESSO. *J. Phys. Condens. Matter* **29**, 465901 (2017).
69. van Setten, M. J. et al. The PseudoDojo: training and grading a 85 element optimized norm-conserving pseudopotential table. *Comput. Phys. Commun.* **226**, 39–54 (2018).
70. Dudarev, S. L., Botton, G. A., Savrasov, S. Y., Humphreys, C. J. & Sutton, A. P. Electron-energy-loss spectra and the structural stability of nickel oxide: an LSDA+U study. *Phys. Rev. B* **57**, 1505–1509 (1998).
71. Cococcioni, M. & de Gironcoli, S. Linear response approach to the calculation of the effective interaction parameters in the LDA+U method. *Phys. Rev. B* **71**, 035105 (2005).
72. Topsakal, M. & Wentzcovitch, R. M. Accurate projected augmented wave (PAW) datasets for rare-earth elements (RE = La–Lu). *Comput. Mater. Sci.* **95**, 263–270 (2014).
73. Dal Corso, A. Pseudopotentials periodic table: from H to Pu. *Comput. Mater. Sci.* **95**, 337–350 (2014).
74. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
75. Peterman, D. J., Weaver, J. H. & Peterson, D. T. Electronic structure of metal hydrides. V. *x*-dependent properties of LaH_x (1.9 < ~*x* < 2.9) and NdH_x (2.01 < ~*x* < ~2.27). *Phys. Rev. B* **23**, 3903–3913 (1981).
76. Knappe, P., Müller, H. & Mayer, H. W. Tetragonal rare earth hydrides $REH(D)_{\approx 2.33}$ (RE = La, Ce, Pr, Nd, Sm) and a neutron diffraction study of $NdD_{2.36}$. *J. Less Common Metals* **95**, 323–333 (1983).
77. Errea, I. et al. Quantum hydrogen-bond symmetrization in the superconducting hydrogen sulfide system. *Nature* **532**, 81–84 (2016).

Acknowledgements We thank B. Brennessel from the Department of Chemistry at the University of Rochester for providing the technical assistance during the XRD and elemental analysis. We thank M. Debessai for his assistance on the coil setup for the magnetic-susceptibility measurements. Also, we thank I. Silvera and I. Hogarth for the useful scientific discussions and R. C. Heist and L. Koelbl for reading through the manuscript and providing valuable suggestions. Preparation of diamond surfaces and EDX measurements were performed in part at the University of Rochester Integrated Nanosystems Center. Computational resources were provided by the Center for Integrated Research Computing at the University of Rochester. This research was supported by NSF grant no. DMR-2046796, Unearthly Materials Inc. and US Department of Energy, Office of Science, Fusion Energy Sciences under award number DE-SC0020340.

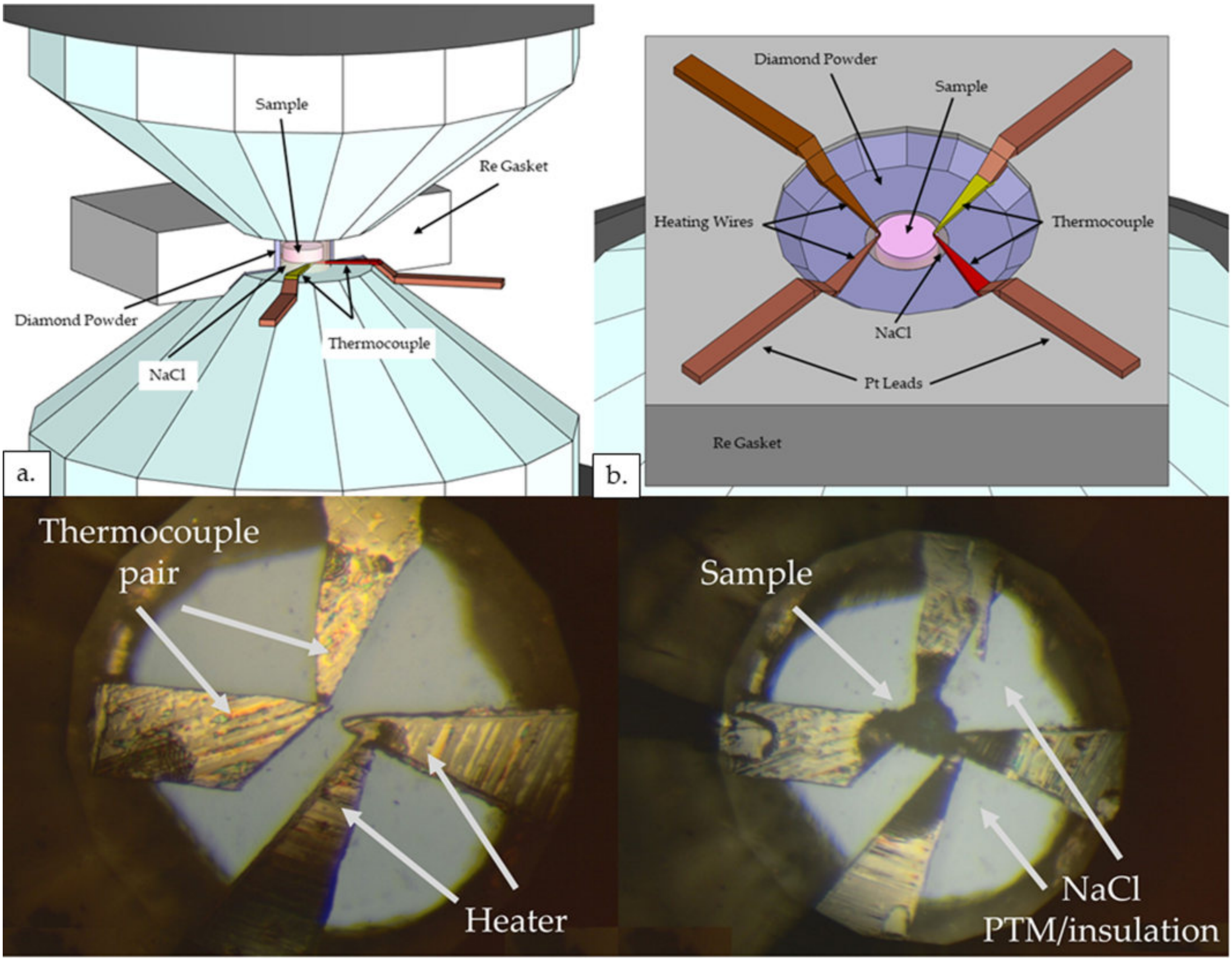
Author contributions N.D.-G., E.S., R.M. and H.P. contributed equally to this work as co-first authors. E.S., D.D., N.D.-G., R.M., H.P. and R.P.D. contributed to performing the electrical-conductivity measurements. N.D.-G., N.K.-S., S.M., S.E.D. and R.P.D. contributed to performing a.c. magnetic-susceptibility measurements and analysed the data. N.D.-G., R.M. and R.P.D. contributed to performing heat-capacity measurements and the analysis. E.S., N.D.-G., R.M., D.D., H.P. and S.E.D. contributed to performing elemental analysis, EDX studies and XRD measurements. H.P., R.M., S.E.D. and R.P.D. contributed to performing Raman studies and H.P. and R.P.D. analysed the data. S.E.D. and A.S. performed structure analysis. H.P., S.E.D. and R.P.D. performed the magnetization measurements using a PPMS and R.P.D. analysed the data. K.V.L. and A.S. performed the simulations and analysed the data and chemistry protocol. N.D.-G., K.V.L., A.S., S.E.D. and R.P.D. wrote the paper. All authors discussed the results and commented on the manuscript. R.P.D. conceived the project and oversaw the entire project.

Competing interests The University of Rochester (U of R) has patents pending related to the discoveries of R.P.D. in the field of superconductivity. R.P.D. is a cofounder and chairman of the board of Unearthly Materials Inc. (UM), a Delaware corporation. UM has licensing agreements with U of R related to the patents, proprietary interests and commercialization rights related to the scientific discoveries of R.P.D. UM, U of R and R.P.D. are subject to non-disclosure and confidentiality agreements. A.S. is a cofounder, president, chief executive officer and board member of UM.

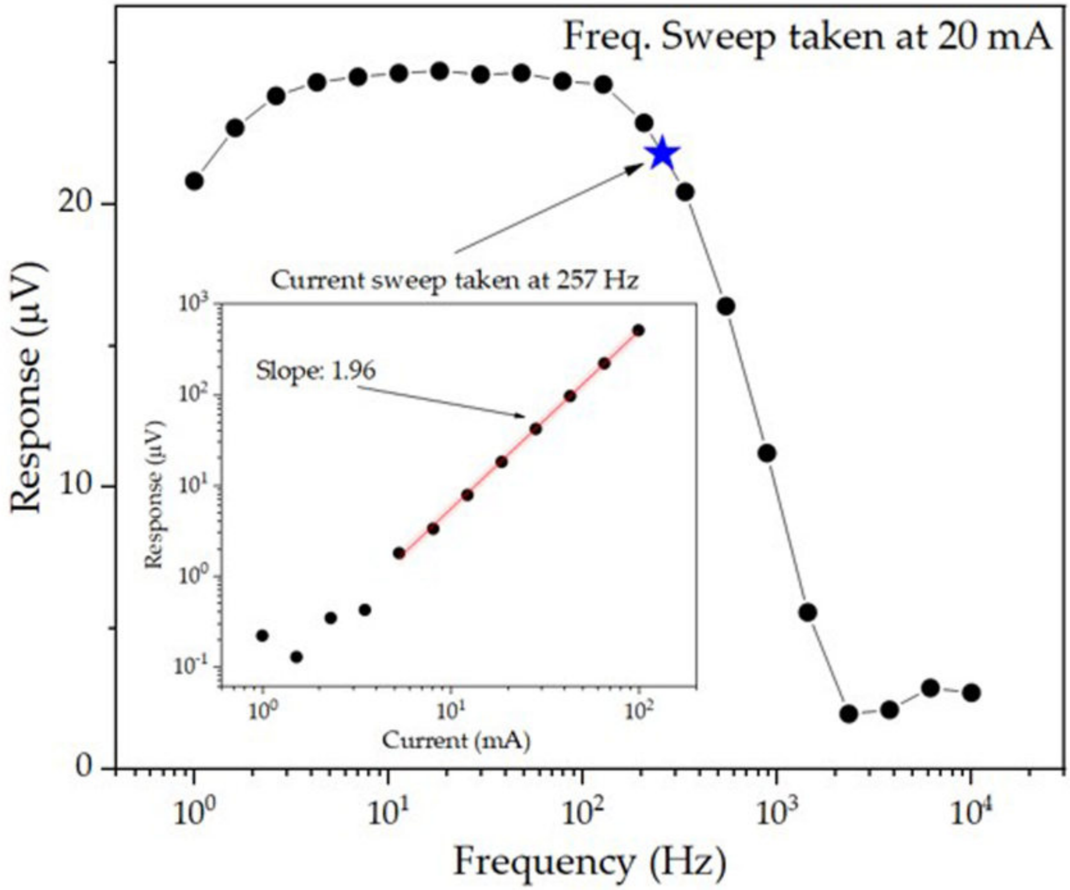
Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-05742-0.
Correspondence and requests for materials should be addressed to Ranga P. Dias.
Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



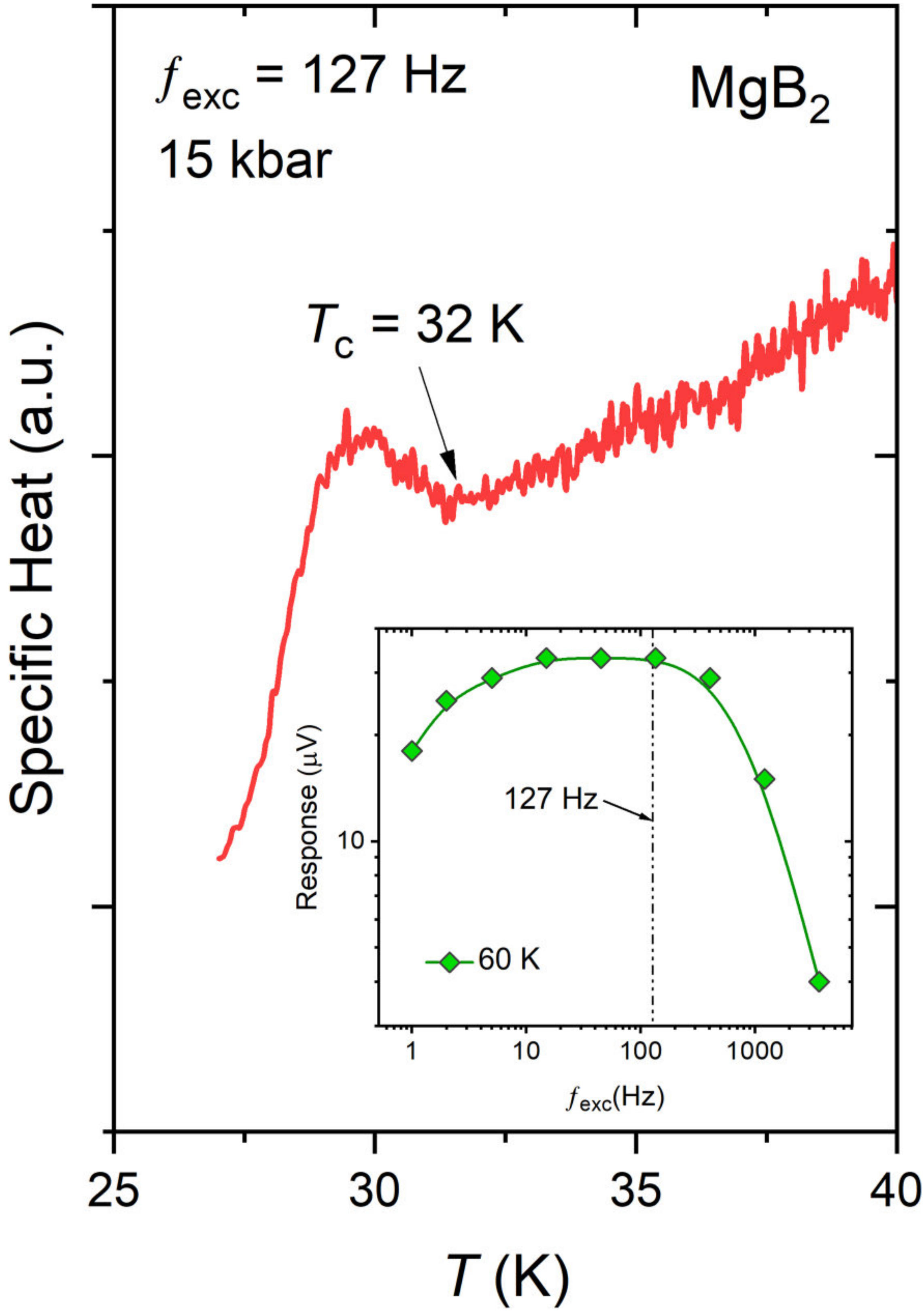
Extended Data Fig. 1 | Raman spectra. a, The spectral deconvolution of Raman spectra of compound A on compression. **b**, The Raman shift versus pressure of compound A at high pressures, indicating the three distinct phases. **c**, The spectral deconvolution of Raman spectra of compound B on compression.



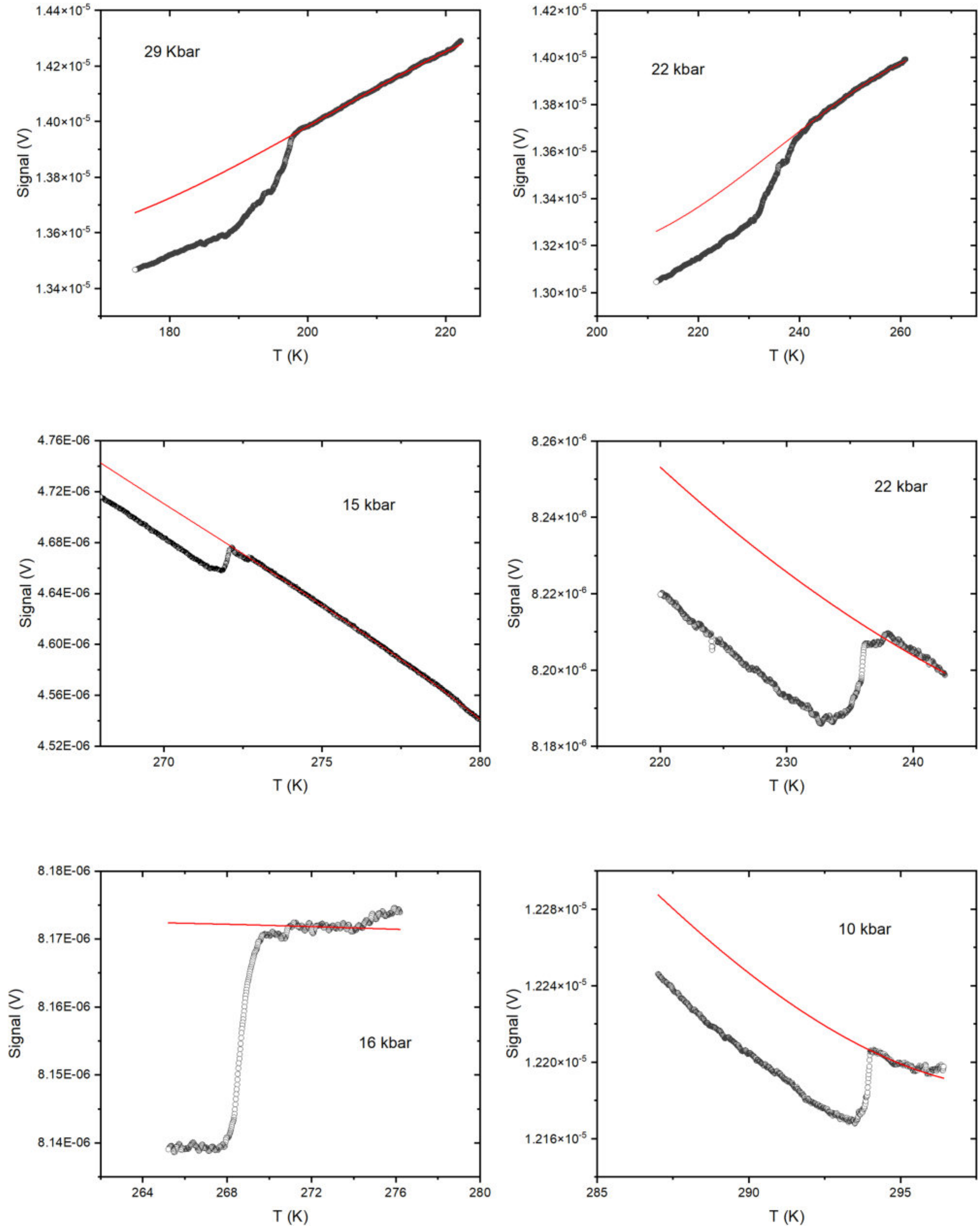
Extended Data Fig. 2 | The heat-capacity setup. Top, schematic rendering of the new a.c calorimetry technique (not to scale). The sample is surrounded by a NaCl insert with a heater and thermocouple making contact with the sample. **a**, View of the preparation as seen from the side showing the thermocouple making contact with the sample inside the DAC. **b**, View of the preparation as seen from the top of the sample area showing the configuration of heater, thermocouple and Pt leads. Bottom left, heat-capacity setup before loading sample. The thermocouple consists of a shorted alumel/chromel pair. The heater pair consists of a shorted metal, nichrome, Ti or Pt. When driven at frequency *f*, the sample temperature modulates at frequency $2 \times f$, which manifests as a voltage on the thermocouple pair that can be measured by a lock-in amplifier. Bottom right, after the sample is loaded, in contact with both the heater and the thermocouple, a small piece of NaCl is placed on top to thermally insulate it from the diamond.



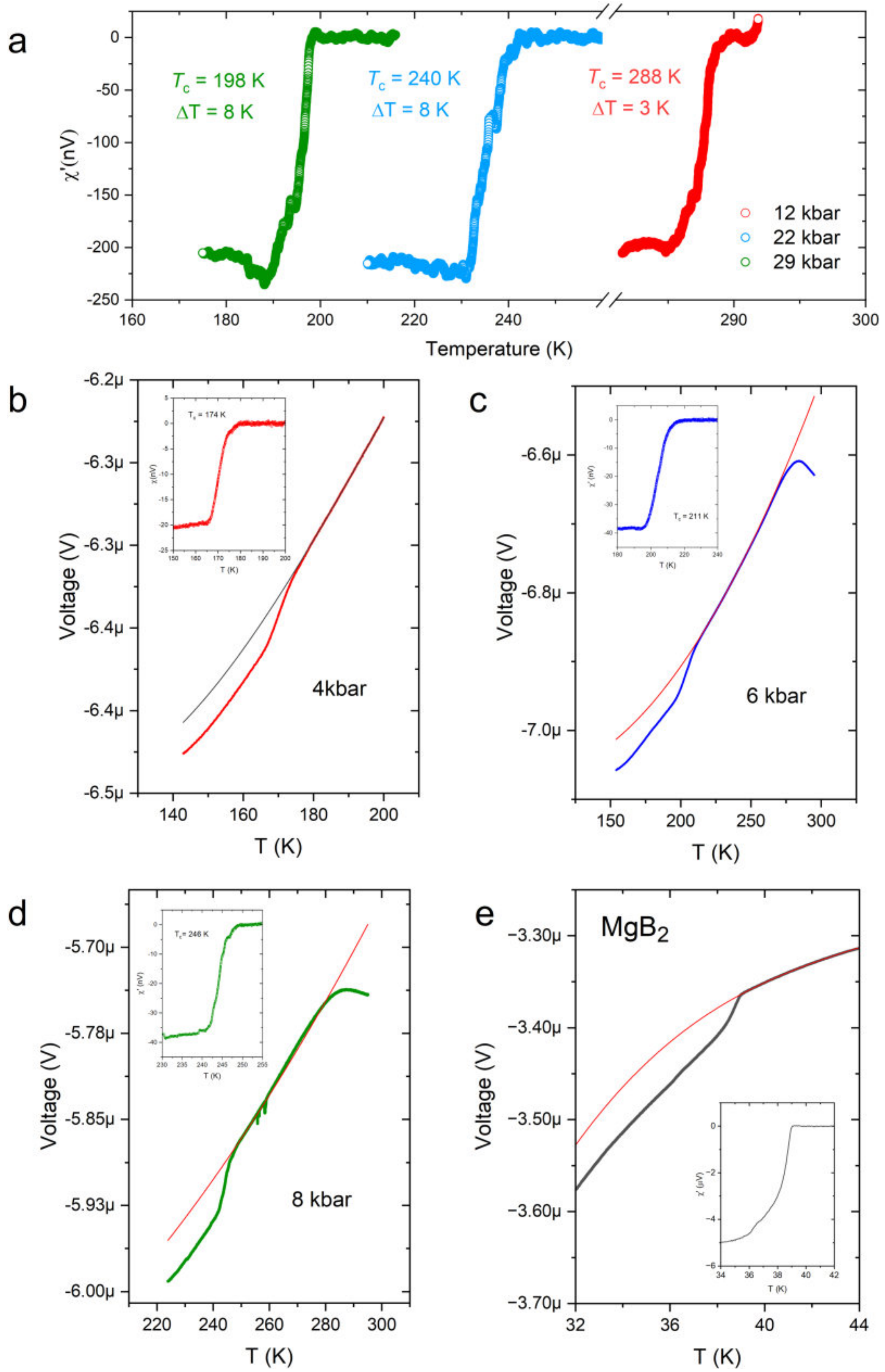
Extended Data Fig. 3 | Frequency response. Frequency and current sweeps measured on a heat-capacity setup before running the experiment. The frequency sweep shows the characteristic plateau and the current sweep demonstrates quadratic dependence, as expected from ohmic heating.



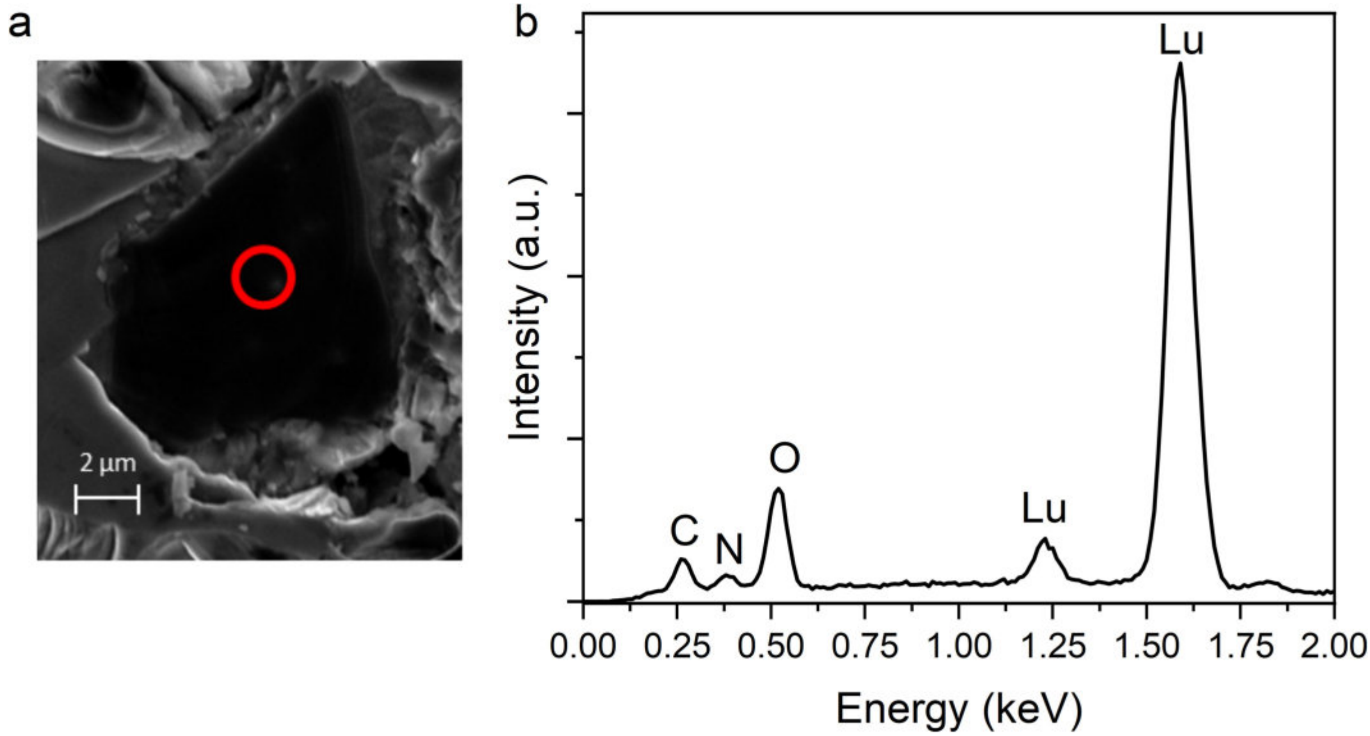
Extended Data Fig. 4 | Heat capacity. Specific heat capacity of MgB_2 as a function of temperature at 15 kbar and 127 Hz. The superconducting signature is clearly observed at 32 K. Inset, recorded lock-in voltages during the frequency sweeps at 60 K.



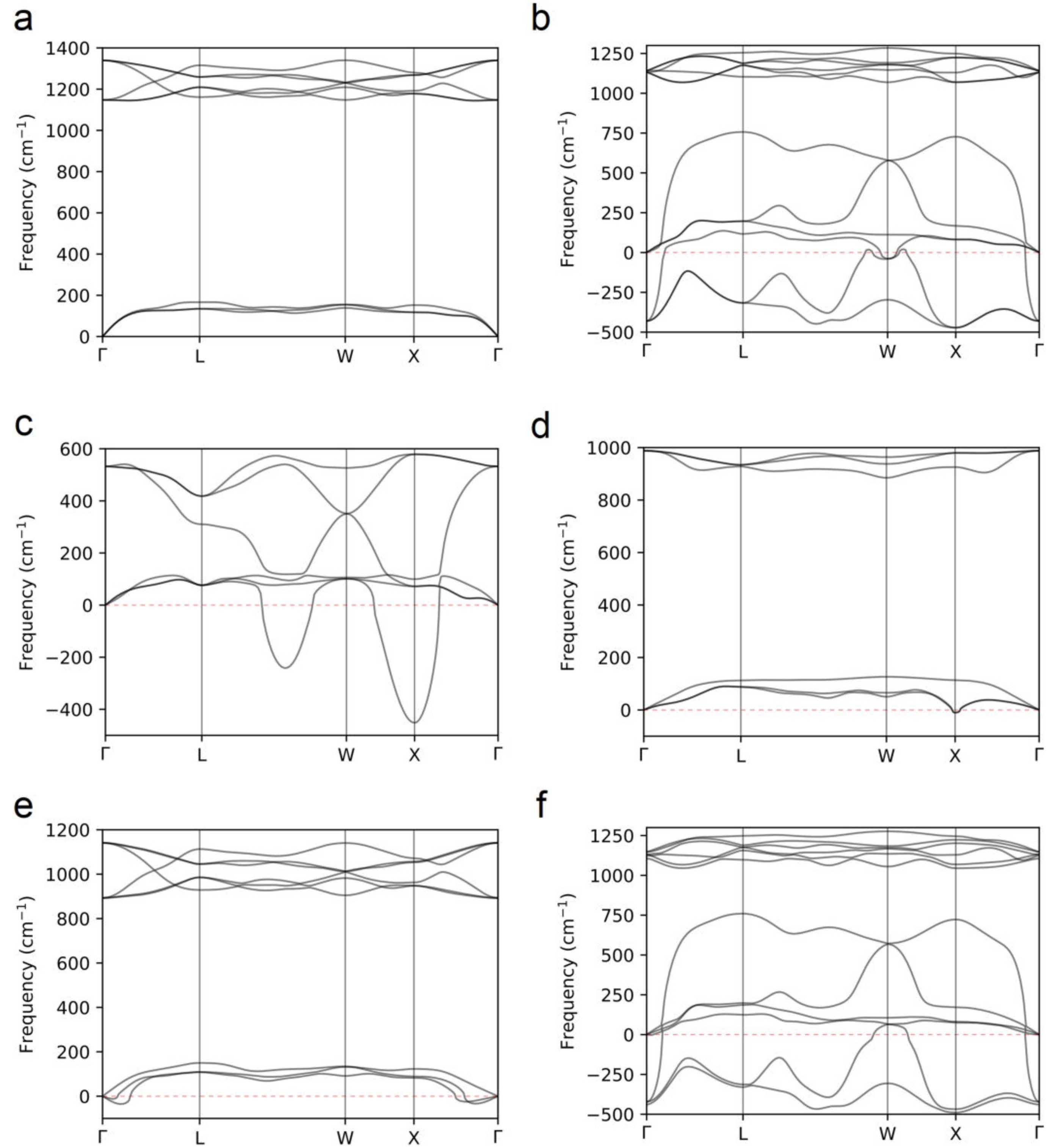
Extended Data Fig. 5 | a.c. susceptibility data before background subtraction. Voltage in volts versus temperature plots at different pressures before the background subtraction. Cubic or quadratic polynomial background was used for background subtraction for susceptibility data. This figure shows fittings with cubic or quadratic polynomials indicated by the red lines. For a.c. susceptibility data, the background subtraction was done mainly for visualization purposes.



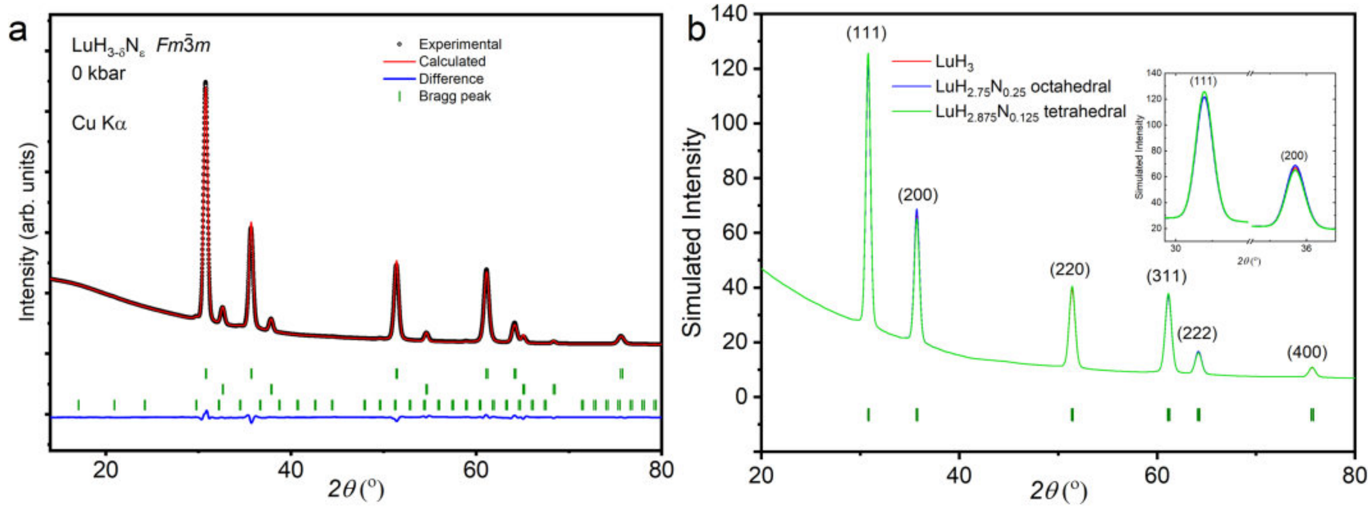
Extended Data Fig. 6 | Further a.c. susceptibility measurements. a, The a.c. susceptibility in nanovolts versus temperature for a larger sample of the N-doped Lu hydride system at select pressures from run 2, showing large diamagnetic signal of the superconducting transition owing to the large volume of the sample. The superconducting transition shifts rapidly under pressure to lower temperatures. a.c. susceptibility measurements taken over broader temperature ranges for N-doped Lu hydride at 4 kbar (**b**), 6 kbar (**c**) and 8 kbar (**d**). The red line in **b**–**d** is the quadratic fit for the background and the insets show the signals with the background subtracted. **e**, a.c. susceptibility measurements of MgB_2 as a function of temperature using exact same coil set up as the test sample.



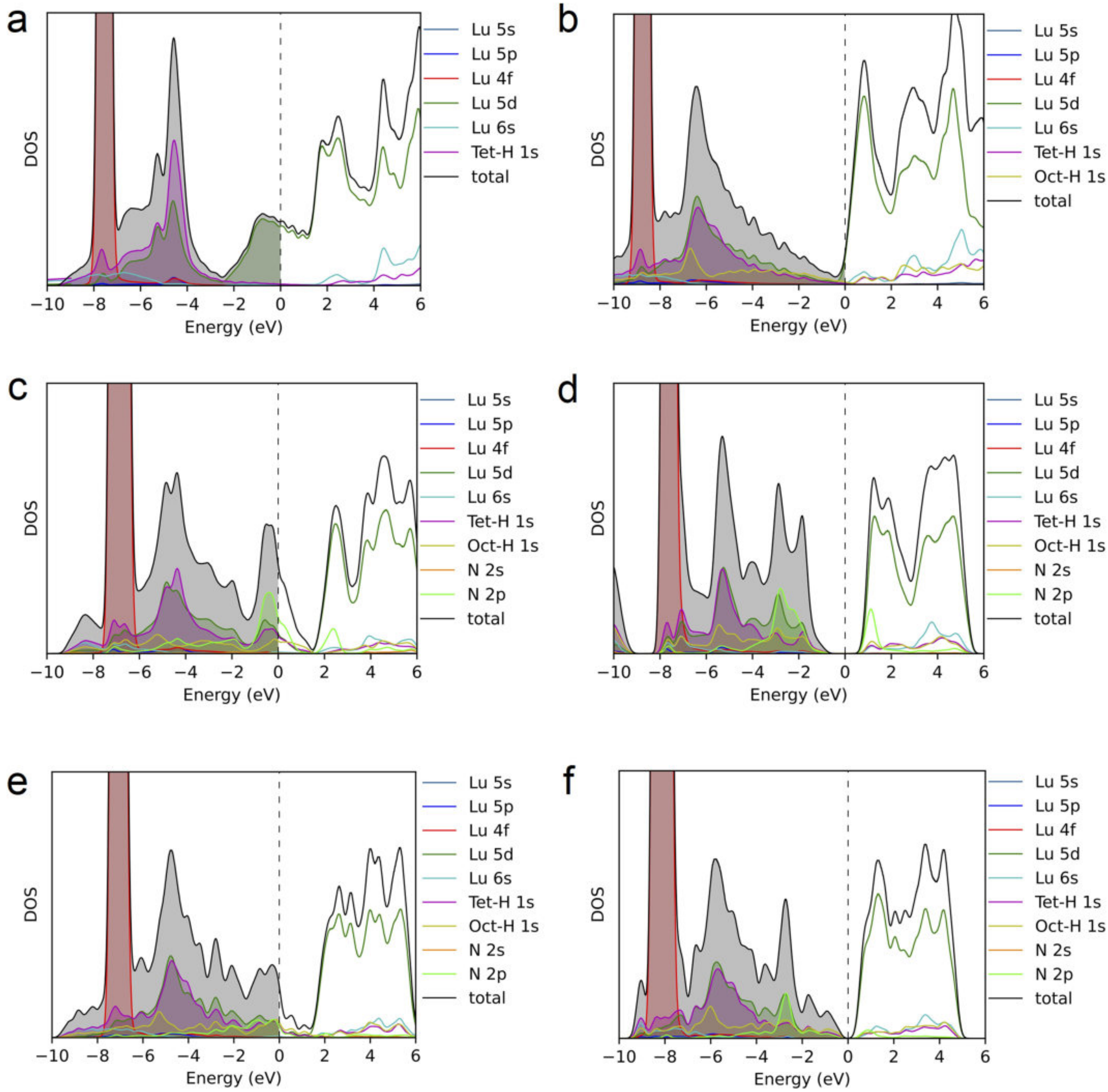
Extended Data Fig. 7 | EDX measurements. For EDX measurements, samples were prepared by mounting on an aluminium pin mount with double-sided carbon tape. The samples were then imaged using a Zeiss Auriga scanning electron microscope. Regions of interest were chosen by comparing the scanning electron microscopy image to a white-light image taken beforehand. EDX measurements were performed in the Zeiss Auriga scanning electron microscope with a driving energy of 15 kV and collected and analysed using an EDAX detector with the EDAX APEX software. Carbon and aluminium peaks seen in the EDX spectra originating from the carbon tape and aluminium mount required to place the samples into the scanning electron microscope vacuum chamber. EDX measurements provide further evidence for the presence of nitrogen in our samples.



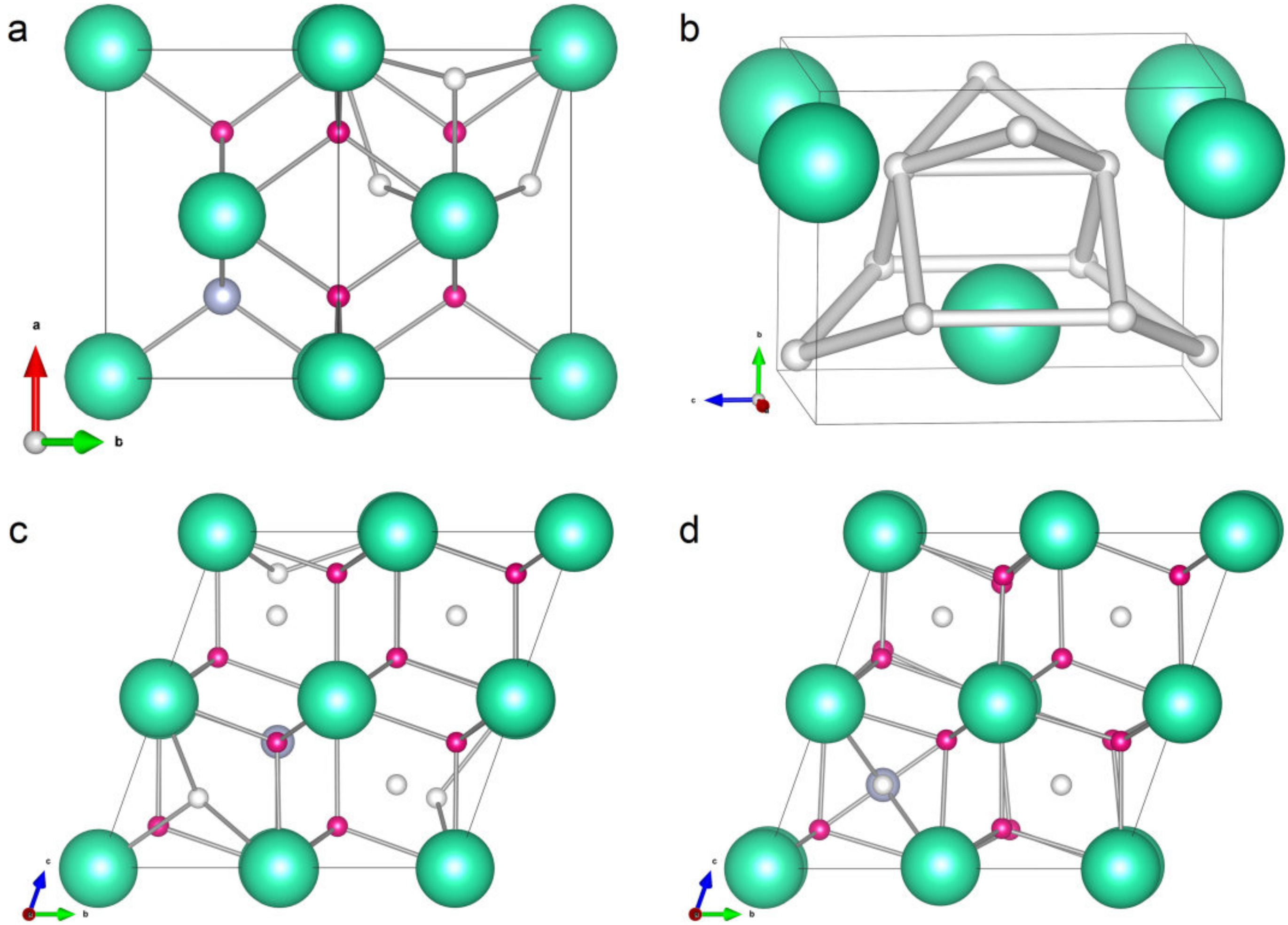
Extended Data Fig. 8 | Phonon bands of stoichiometric Lu hydrides. The calculated phonon band structures of 0 kbar LuH_2 in the fluorite structure (**a**), $Fm\bar{3}m$ LuH_3 (**b**), LuH in the RS structure (**c**) and LuH in the ZB structure (**d**). **e**, The calculated phonon band structures of 0 kbar LuH_2 in the fluorite structure using a triclinic representation of the lattice vectors with x parallel to a and z parallel to c^*, as opposed to the more highly symmetric lattice vectors for a primitive cell of a fcc cell; in this representation, the structure is represented with D_{3d} point-group symmetry as opposed to O_h point-group symmetry as in **a**. **f**, The calculated phonon band structures of 0 kbar LuH_3 using the same triclinic representation of the lattice vectors and point-group symmetry as in **e**.



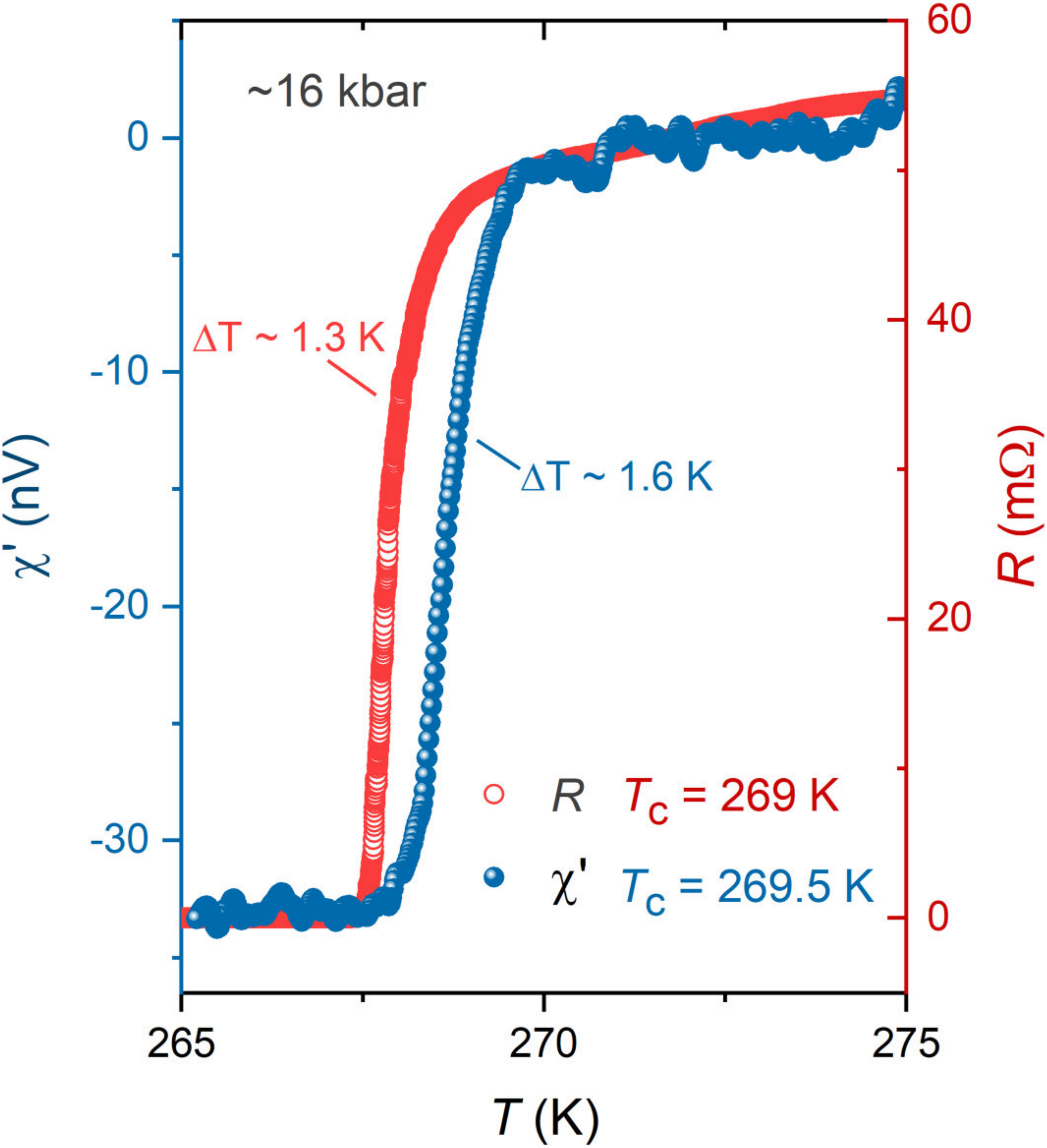
Extended Data Fig. 9 | Rietveld refinement of site occupancies. a, Rietveld refinement of the X-ray powder diffraction data collected at 295 K with Cu Kα radiation with refining the occupancy of the tetrahedral interstitial site with N for nitrogen-doped lutetium hydride. **b**, Simulation of the XRD pattern with Cu Kα wavelength for LuH_3 (red), LuH_3 with a N replacing a single H in an octahedral site (blue) and a tetrahedral site (green). Rietveld refinement of the X-ray powder diffraction data of ground powder sample was performed with an attempt to investigate the possible N substitution in nitrogen-doped lutetium hydride. We note here that XRD is mostly dominated by heavy Lu atoms.



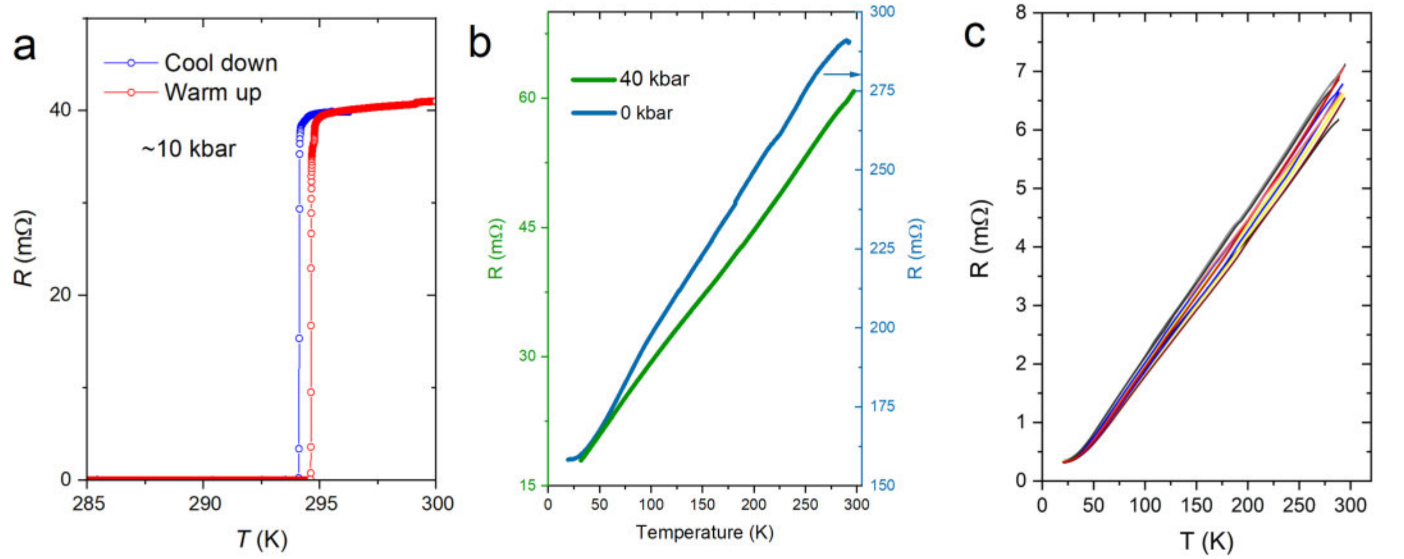
Extended Data Fig. 10 | Projected density of states. The atom and angular momentum projected partial density of states of LuH_2 in the fluorite structure (**a**); $Fm\bar{3}m$ LuH_3 (**b**); the cubic cell of $Fm\bar{3}m$ LuH_3 with a N substituted for a H in an octahedral (**c**) and tetrahedral (**d**) interstice; and a 2 × 2 × 2 supercell of the rhombohedral primitive cell of $Fm\bar{3}m$ LuH_3 with a N substituted for a H in an octahedral (**e**) and tetrahedral (**f**) interstice. In the legends, Oct- means hydrogens in the octahedral interstices and Tet- means hydrogens in the tetrahedral interstices. Each channel is summed over all similar atoms in the unit cell and the plots are scaled to represent a maximum value of 2.5 states eV^{-1} per formula unit.



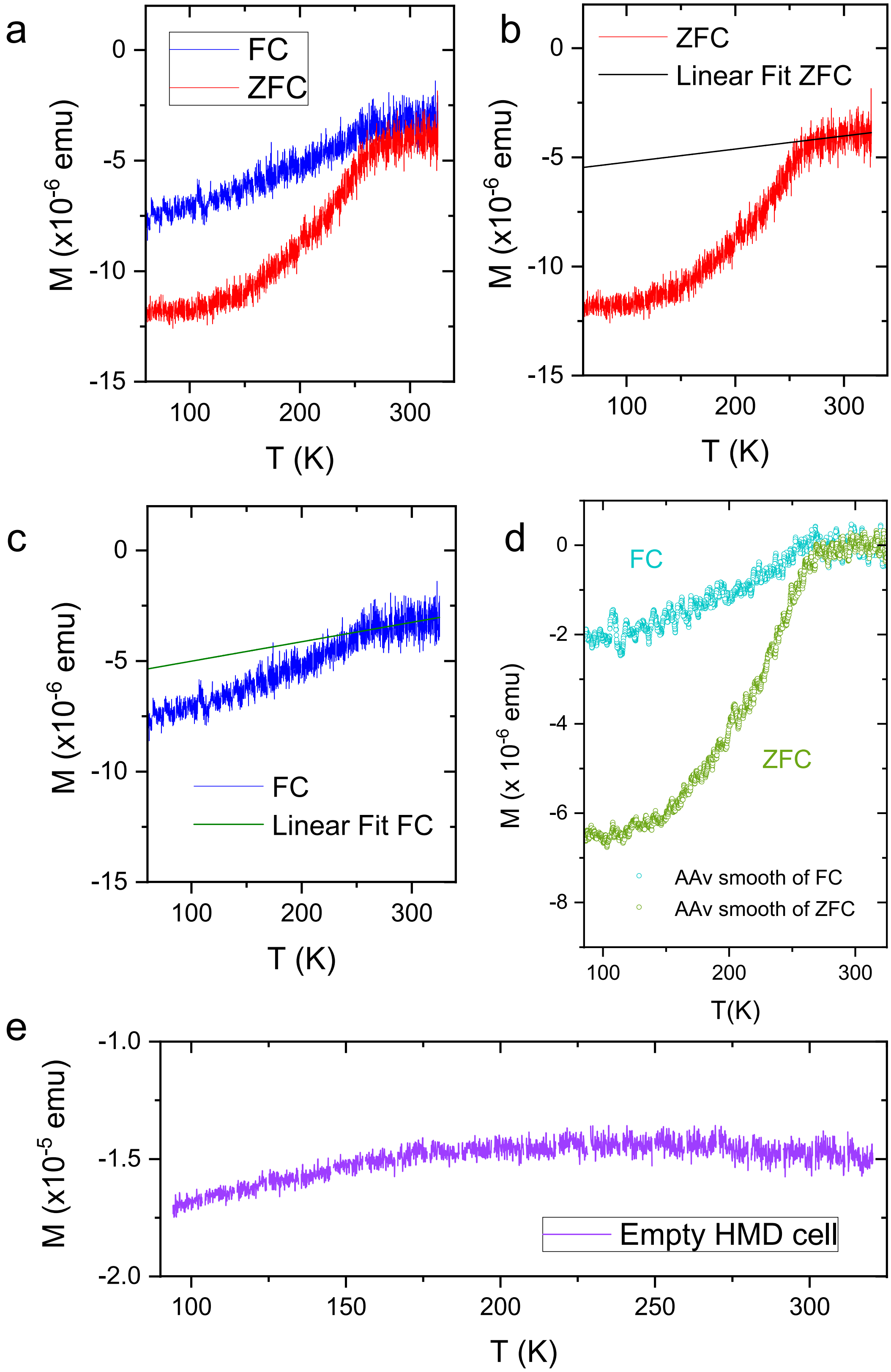
Extended Data Fig. 11 | Distorted structures predicted by DFT. **a**, The distortions to the octahedral hydrogens observed by substituting a N atom for a tetrahedral atom in a single unit cell of *Fm*$\bar{3}$*m* LuH_3. **b**, The *Pmnm* LuH_3 structure found by perturbing the cubic *Fm*$\bar{3}$*m* unit cell of LuH_3, which suggests possible light-atom positions in phase III. **c**, The lattice distortions from substituting a N into a tetrahedral interstice in a 2 × 2 × 2 supercell of the rhombohedral primitive of LuH_3. **d**, The lattice distortions from substituting a N into an octahedral interstice in a 2 × 2 × 2 supercell of the rhombohedral primitive of LuH_3. The lutetium atoms are green, the nitrogen atoms are lavender and the hydrogen atoms in octahedral interstitial sites are white and those in tetrahedral interstitial sites are pink. In **b**, there is no distinction made between the hydrogen atom sites, so they are all white.



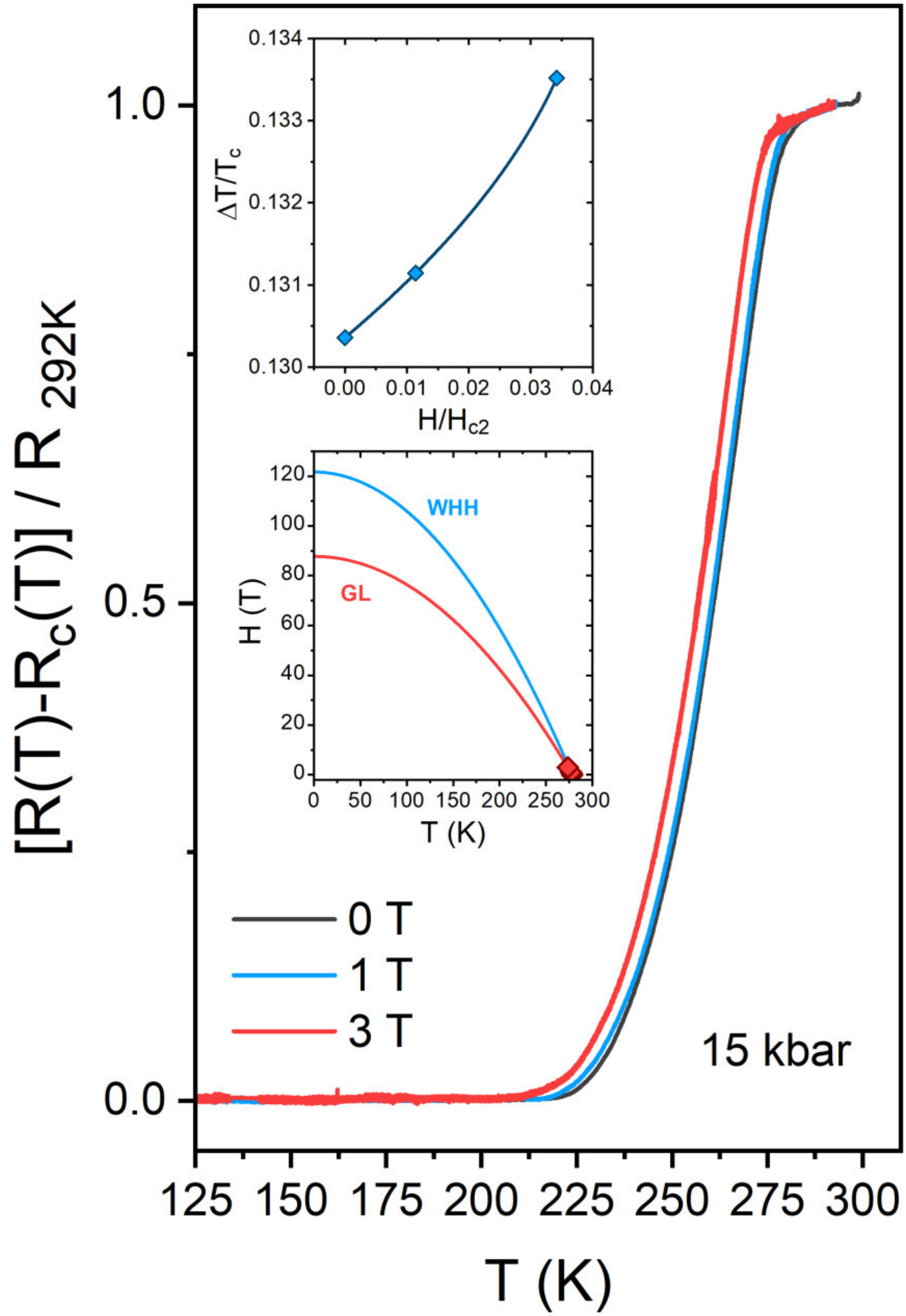
Extended Data Fig. 12 | Superconducting transition widths. For comparison, the superconducting transition obtained from electrical measurements and a.c. susceptibility measurement at a similar pressure (16 kbar) is shown by red and blue, respectively. The transition width of the resistance drop is 1.3 K and 1.6 K for the a.c. magnetic susceptibility measurement.



Extended Data Fig. 13 | Low-temperature electrical-resistance behaviour of N-doped Lu–H systems. **a**, The resistance measured on both warming and cooling at about 10 kbar. **b**, Temperature-dependent electrical resistance of phases I and III, showing the non-superconducting state. **c**, Four-probe electrical-resistance measurements of different Lu–H–N samples, which consistently shows highly metallic behaviour with decreasing temperature.

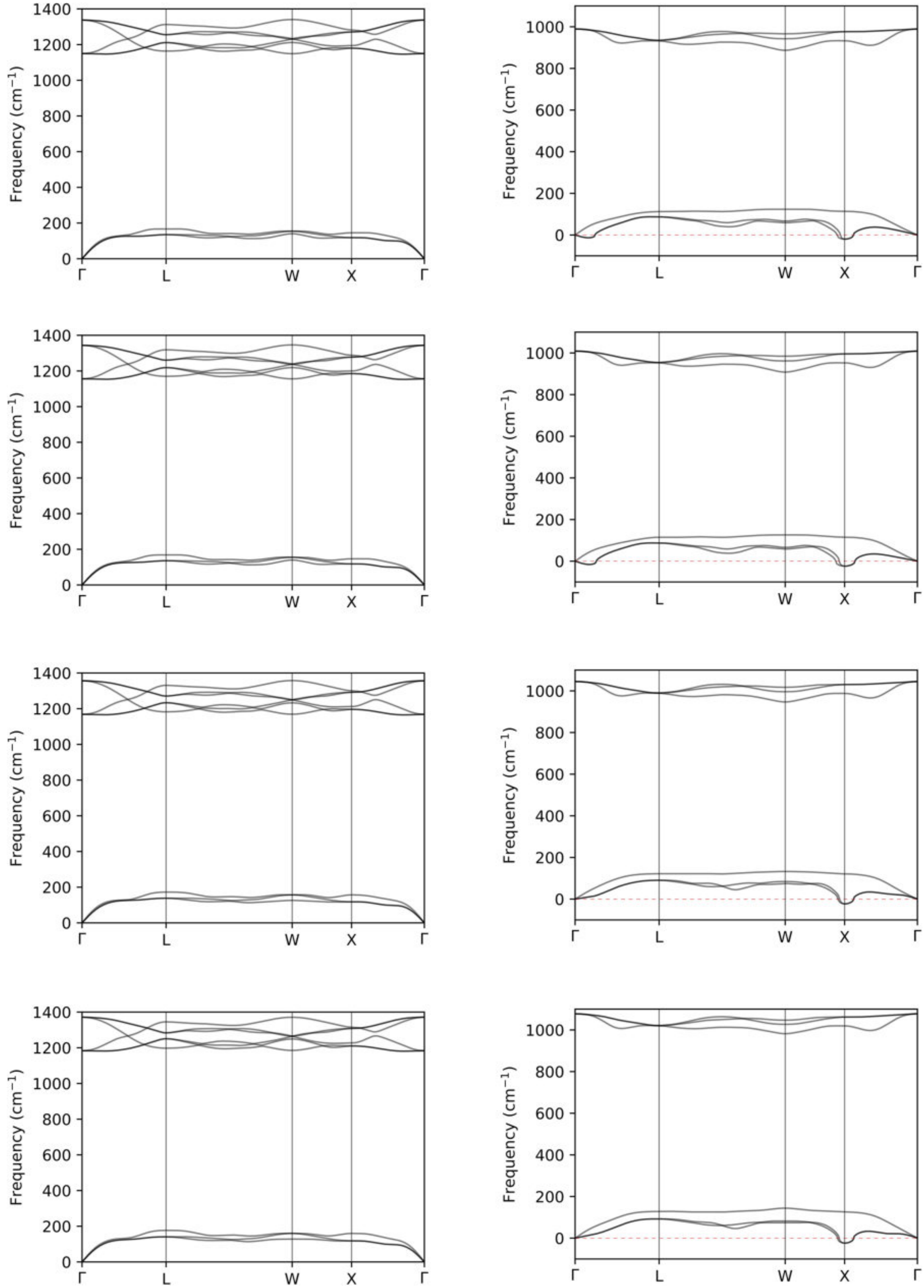


Extended Data Fig. 14 | Magnetic-susceptibility background and smoothing. a–**c**, The ZFC and FC magnetization versus temperature at 8 kbar used to construct Fig. 3a, along with a linear fit to the data at temperatures above the transition temperature, which was used for the background subtraction. **d**, The ZFC and FC curves with the linear backgrounds shown in **b** and **c** subtracted out, as well as with a ten-point adjacent-average smoothing applied. **e**, The measured cell background at 60 Oe for the HMD cell used for the d.c. measurements.



Extended Data Fig. 15 | See next page for caption.

Extended Data Fig. 15 | Electrical-resistance behaviour under magnetic field. Low-temperature electrical-resistance behaviour under magnetic fields of H = 0 T, 1 T and 3 T (increasing from right to left) at 15 kbar. In this study, the superconductivity of nitrogen-doped lutetium hydride is suppressed by the application of a 3-T external magnetic field, reducing T_c by about 5 K at 15 kbar, further confirming a superconducting transition. The temperature dependence of the resistance of a simple metal is written as: $R(T) = R_o + aT^2 + bT^5$. We fit the data below T < 220 K for each field, at which the resistance goes to the minimum value, to that function and subtracted it out. Inset top, the superconducting transition width, ΔT_c, at 15 kbar slightly increases under external magnetic fields. The ΔT_c has a good linear relationship with the applied magnetic field, as expected from the percolation model. The superconducting transition width is defined here as $\Delta T_c = T_{90\%} - T_{10\%}$, in which $T_{90\%}$ and $T_{10\%}$ are the temperatures corresponding to 90% and 10% of the resistance at 292 K, respectively. Fitting to the linear relation of $\Delta T_c = \Delta T_c(0) + kH_{c2}$, in which $\Delta T_c(0)$ is the width at zero external field and k is a constant, provides the values $\Delta T_c(0)$ = 36.3 K and k = 0.07 KT^{-1}. The large transition width at zero field indicates sample inhomogeneities, which is typical for high-pressure experiments. Inset bottom, the temperature dependence of the upper critical field, $H_c(T) = H_c(0)\left[1 - \left(\frac{T}{T_c}\right)^2\right]$, can be expressed using GL theory or the conventional Werthamer–Helfand–Hohenberg model. The GL model in the limit of zero temperature yields $H_{c2}(0) \approx 88$ T. From the Werthamer–Helfand–Hohenberg model in the dirty limit, $H_{c2}(0)$ can be extrapolated from the slope of the H–T curve as $H_{c2}(0) = 0.693 \left|\frac{dH_{c2}}{dT}\right|_{T=T_c} T_c$, which yields roughly 122 T.



Extended Data Fig. 16 | Phonon bands of pressurized stoichiometric Lu hydrides. The calculated phonon band structures of LuH_2 in the fluorite structure (left) and LuH in the ZB structure (right) at 0 kbar (top row), 10 kbar (second row), 30 kbar (third row) and 50 kbar (bottom row). The electronic smearing width is 0.005 Ry and the lattice vectors are the highly symmetric ones for a fcc cell. Negligible change in the computed electron–phonon couplings or logarithmic frequency is seen for LuH_2 on pressurization.

Article

Population-based heteropolymer design to mimic protein mixtures

https://doi.org/10.1038/s41586-022-05675-0

Received: 13 June 2022

Accepted: 21 December 2022

Published online: 8 March 2023

Check for updates

Zhiyuan Ruan[1], Shuni Li[2], Alexandra Grigoropoulos[1], Hossein Amiri[3], Shayna L. Hilburg[4], Haotian Chen[1], Ivan Jayapurna[1], Tao Jiang[1,11], Zhaoyi Gu[1,12], Alfredo Alexander-Katz[4], Carlos Bustamante[3,5,6,7,8], Haiyan Huang[2,9] & Ting Xu[1,6,10] ✉

Biological fluids, the most complex blends, have compositions that constantly vary and cannot be molecularly defined[1]. Despite these uncertainties, proteins fluctuate, fold, function and evolve as programmed[2–4]. We propose that in addition to the known monomeric sequence requirements, protein sequences encode multi-pair interactions at the segmental level to navigate random encounters[5,6]; synthetic heteropolymers capable of emulating such interactions can replicate how proteins behave in biological fluids individually and collectively. Here, we extracted the chemical characteristics and sequential arrangement along a protein chain at the segmental level from natural protein libraries and used the information to design heteropolymer ensembles as mixtures of disordered, partially folded and folded proteins. For each heteropolymer ensemble, the level of segmental similarity to that of natural proteins determines its ability to replicate many functions of biological fluids including assisting protein folding during translation, preserving the viability of fetal bovine serum without refrigeration, enhancing the thermal stability of proteins and behaving like synthetic cytosol under biologically relevant conditions. Molecular studies further translated protein sequence information at the segmental level into intermolecular interactions with a defined range, degree of diversity and temporal and spatial availability. This framework provides valuable guiding principles to synthetically realize protein properties, engineer bio/abiotic hybrid materials and, ultimately, realize matter-to-life transformations.

Whereas molecular precision can be readily achieved inside test tubes, biological fluids are diverse, complex and full of uncertainty[7]. Evolutionarily, the selection of the fittest proteins depends on their surroundings. As a result, natural proteins harmoniously coexist with each other and collectively execute complex tasks with exceptional fidelity amid random fluctuations and external perturbations (Fig. 1a)[8–10]. Information embedded in the sequence space of natural proteins provides the blueprint to design synthetic heteropolymers[11,12] to achieve predictable interplay with biological components as protein substitutes, to holistically recapitulate collective behaviours in protein mixtures and to further gain functions of special proteins while maintaining system compatibility.

We propose that the chemical and sequence characteristics of proteins at the segmental level, as opposed to the monomeric level, is the key factor governing how they transiently interact with neighbouring molecules and the collective behaviours of biological fluids. Experimentally, mimicking the length distribution of blocks containing consecutive residues of the same characteristics in soluble proteins has been proven effective in designing random heteropolymers (RHPs) to stabilize proteins in non-aqueous media as well as to mimic channel proteins to transport protons[13–15]. In both cases, the RHP chains have substantially reduced conformational freedom, being either anchored at polar or non-polar interfaces or spanning a lipid bilayer; and the observed function readouts are from a subpopulation within each designed RHP ensemble. Nevertheless, these results have validated the importance and value of extracting protein sequence information beyond the monomeric level.

In biological fluids, components often interact through random encounters and the whole population needs to be considered. However, when the block-length-based analysis was extended beyond soluble proteins, the block-length distribution broadened (Extended Data Fig. 1 and Supplementary Figs. 1 and 2). Furthermore, the results regarding the block-length distribution contain no information regarding their sequential arrangement within each protein chain. Yet, the effective hydrophobicity of a given block depends on the chemical characteristics of neighbouring ones[14], which affects the probability of it being

[1]Department of Materials Science and Engineering, University of California Berkeley, Berkeley, CA, USA. [2]Department of Statistics, University of California Berkeley, Berkeley, CA, USA. [3]Institute for Quantitative Biosciences-QB3, University of California, Berkeley, CA, USA. [4]Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [5]Department of Molecular and Cell Biology, University of California Berkeley, Berkeley, CA, USA. [6]Department of Chemistry, University of California Berkeley, Berkeley, CA, USA. [7]Department of Physics, University of California Berkeley, Berkeley, CA, USA. [8]Howard Hughes Medical Institute, University of California Berkeley, Berkeley, CA, USA. [9]Center for Computational Biology, University of California, Berkeley, CA, USA. [10]Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [11]Present address: Department of Chemistry, Xiamen University and The MOE Key Laboratory of Spectrochemical Analysis and Instrumentation, Xiamen, China. [12]Present address: Departments of Chemistry and Biomedical Engineering, Northwestern University, Evanston, IL, USA. ✉e-mail: tingxu@berkeley.edu

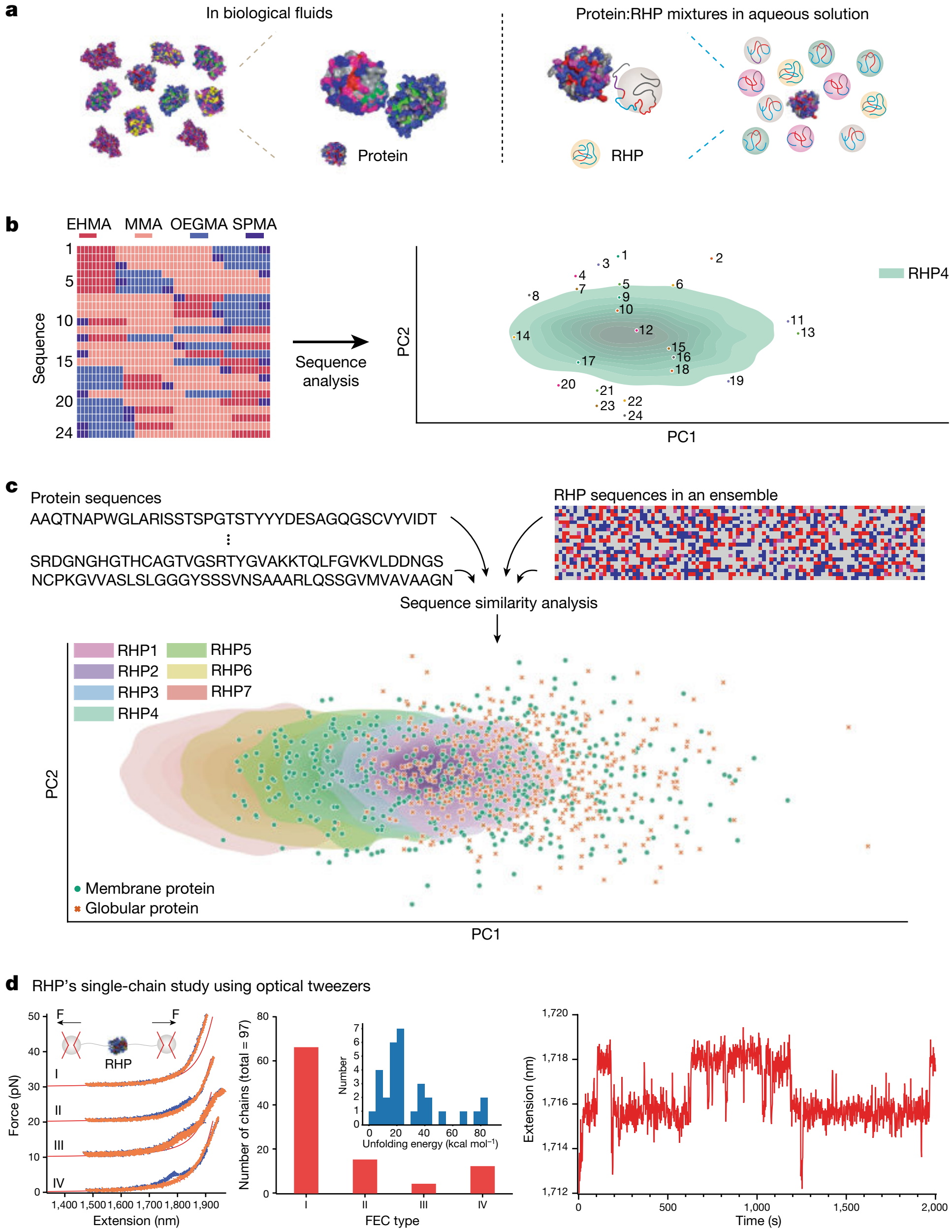


Fig. 1 | Population-based design to mimic the native environments of proteins. a, Heteropolymer ensembles are designed to mimic protein mixtures in biological fluids through transient interactions with neighbouring proteins. **b**, The left shows the permutated sequences in descending order of PC2 value from top to bottom. The right shows the projection of 24 permutated sequences with RHP4 composition ratio and RHP4 ensemble onto the PCA space. **c**, 2D sequence analysis of proteins and heteropolymers can be used to compare similarities in their chemical characteristics at the segmental level. 100 membrane protein sequences (green dot) and 100 globular protein sequences (orange cross) (both randomly selected) are shown for clarity. Each shaded area is a univariate distribution of an ensemble of 2,000 RHP chains using kernel density estimation. **d**, Single-chain study of RHP4 using optical tweezers. The left panel shows four representative types of FEC during pulling (blue) and relaxing (orange): no unfolding signature (type I), discrete rips (type II), smooth shoulders (type III) and a mixture of rips and shoulders (type IV). The red solid curve represents the extensible worm-like chain model for 6 kb DNA. The plots are shifted along the *y* axis for clarity by a step of +10 pN. The number of RHP4 chains that show each type of FEC and the distribution of unfolding energies for type II–IV chains (inset) are shown in the middle panel. The right panel shows the passive-mode trace reflecting the unfolding and refolding transitions of an RHP4 chain with type IV behaviour.

surface exposed as well as its spatial location within a globular RHP chain. At the whole chain level, the sequential arrangement of blocks will affect the overall chain conformation, intra- and interchain interactions. Here, we developed a two-dimensional (2D) informative sequence analysis to parameterize and visualize both chemical characteristics and sequential arrangement at the segmental level using an autoencoder model (Fig. 1b) and established design rules to modulate RHPs' similarity to proteins as an ensemble. As a test case, we developed a new library of RHPs as synthetic cytosols capable of modulating RHP–protein and RHP–DNA interactions, and their spatial compartmentalization on microscopic liquid–liquid phase separation (LLPS).

2D informative sequence analysis

The model was trained on a library of protein sequences including membrane proteins and globular proteins collected from the

UniprotKB database[16]. Each sequence was first reduced into four types of pseudo-residue: hydrophilic, hydrophobic, very hydrophobic and charged, on the basis of the respective side-chain hydrophobicity of amino acids (Extended Data Table 1)[17]. Initial sequence analysis was also tested using two or eight pseudo-residues (Supplementary Tables 1 and 2). The choice of four pseudo-residues balances the synthetic feasibility of heteropolymers with the accessible diversity and accuracy of segmental interactions (Supplementary Figs. 3–5)[13]. For each protein, the reassigned protein sequence was truncated into a collection of 50-mer segments for analysis (Extended Data Fig. 1). The 50-residue interval was chosen because it captured most short- and long-range residue–residue contacts in native protein conformation. The residue order and local hydrophobicity of each 50-mer were extracted and represented by a low-dimensional vector, which was projected onto a space comprising the first two principal components (PC1 and PC2) from a principal component analysis (PCA)[18,19]. To a first approximation, the PC1 correlates with the apparent hydrophobicity of the 50-mer; the PC2 correlates with the sequential arrangement of different blocks within the 50-mer. A library of hypothetical chains containing the same four blocks arranged in different orders was analysed for illustration. They can be clearly differentiated on the basis of the 2D sequence analysis (Fig. 1b). The PC1's dependence on the PC2 further highlights the importance of considering the block sequence along a chain and its potential to modulate interchain interactions.

The PCA analysis results for known proteins are shown in Fig. 1c and effectively capture key sequence characteristics for both protein families. Membrane proteins have a subpopulation shifted along the PC1 axis compared to that of globular proteins. Thus, the model captures the distinction between the two protein families through their 50-mer segmental hydrophobicity. The large spatial overlap between the two protein families reflects that common protein motifs are present in both protein families. The two protein families were indistinguishable along the PC2 axis and both are more concentrated in the centre region of PC2. It is reasonable to speculate that proteins with alternating short hydrophobic and hydrophilic segments are less prone to aggregation and are more preferable through evolution.

Design and synthesis of RHP ensembles

Individual heteropolymers cannot fully capture the protein sequence space, and the exact composition of any bio-fluids remains elusive. Thus, a population-based design is required (Supplementary Fig. 6). Individual RHP chains are statistically sequence-defined. They have different monomeric sequences but are statistically similar at the segment level, making them an ideal platform to materialize the segmental information extracted from primary protein sequences[14,20]. A library of RHP ensembles and two diblock heteropolymers (DHP) were designed and synthesized to match the pseudo-residues used in the initial protein sequence analysis (detailed residue assignments are listed in Extended Data Table 1 and Supplementary Tables 1 and 2), using 2–4 out of six selected monomers including methyl methacrylate (MMA), 2-ethylhexyl methacrylate (2-EHMA), 3-sulfopropyl methacrylate potassium salt (3-SPMA), 2-(dimethylamino) ethyl methacrylate (DMAEMA) and oligo (ethylene glycol) methacrylate (OEGMA) with number-average molecular weight (M_n) of 300 or 500 Da, respectively. The MMA:EHMA molar ratio was varied in 5 or 10% increments to modulate the distribution of segmental hydrophobicity. Details of all RHPs studied are shown in Table 1.

For each RHP ensemble, 2,000 simulated RHP sequences based on the Mayo–Lewis equation using experimentally determined reactivity ratios and monomer compositions were analysed[20]. This sample size is representative of the ensemble, because the occupied PCA space was observed to converge at roughly 1,500 chains (Supplementary Fig. 6). The segment distribution from proteins and RHP1–7 was projected onto the same PCA space for similarity comparison (Fig. 1c). The occupied PCA space shifts along the PC1 axis, with some overlapping regions, as the monomer composition varies. RHP ensembles populating the left of the PCA space (lower PC1 value) have more hydrophobic segments than those to the right (Supplementary Fig. 7). The PC2 distribution of RHP ensembles are similar to each other and comparable to those of both protein families. RHP ensembles can be designed to match the segmental diversity of both protein families. There is a sizeable overlap between four RHP ensembles, RHP1–4, with globular proteins; RHP1–7 have some overlap with membrane proteins.

Probing RHP single-chain conformation

Each RHP ensemble covers a defined range of segmental hydrophobicity and their sequential arrangement. We performed single-chain end-to-end pulling and relaxation studies using optical tweezers[21]. RHP4 shows large overlap with both families of proteins in PCA space and was chosen for single-molecule studies. Results from the first 30 chains are statistically similar to those from a total of 97 chains, indicating that the measured RHP chains should be representative of the ensemble. The RHP ensemble mimics the structural heterogeneity of proteins as commonly seen in biological fluids that contain intrinsically disordered, partially folded and folded states. The force-extension curves (FECs) of 66 out of 97 chains showed no deviation from standard worm-like chain behaviour, typical of non-interacting polymers (Fig. 1d and Supplementary Fig. 8). Four out of 97 chains showed no discrete unfolding 'rip' but a 'shoulder' feature in the force range of roughly 5–10 pN (Supplementary Fig. 9), reflecting a process of rapid, quasi-equilibrium unfolding or refolding of structures that are only marginally stable[22]. Fifteen out of 97 chains showed discrete rips (Supplementary Fig. 10) and 12 out of 97 chains showed a combination of rips and shoulders (Supplementary Fig. 11), indicating cooperative unfolding of mechanically stable structures[22]. When one of the nine RHP chains with a combination of rips and shoulders was subjected to a constant external force (passive mode), a dynamic transition between distinct force-extension states was observed (Fig. 1d). The results indicate that certain RHP subpopulations have reversible folding–unfolding transitions similar to those of proteins[23]. FECs were consistent between several pulling and relaxation cycles of individual RHP chains. Three subpopulations comprising 31 out of 97 chains form structures with stability of 29.0 ± 22.3 kcal mol^{-1} (mean ± s.d.). This value falls well within the energy range associated with reversible conformational changes in proteins (roughly 10–90 kcal mol^{-1})[24]. Together, an RHP ensemble that matches the PCA space of protein mixtures has a defined range of segmental characteristics and can mimic the conformational diversity of proteins.

PCA overlap governs RHP/protein interplay

The overlapping PCA regions between proteins and each RHP ensemble indicates similarity in their segmental chemical characteristics, which defines the range of interactions during random encounters in biological fluids. We propose that to design RHPs as protein mixture mimics, it is more essential to capture the range of intra- and intermolecular interactions in biological fluids rather than replicating their exact compositions, which remain undefined and fluctuating. Thus, we evaluate the RHPs' similarity to proteins using two tests. First, we probe how each RHP ensemble interacts with membrane proteins during folding post-translation using a cell-free expression platform. Second, we measure how the presence of RHP ensembles in fetal bovine serum (FBS), the most commonly used biological fluid, affects FBS's ability to support cell cultures during storage and thermal denaturation[25].

Three representative membrane proteins, outer membrane protease (OmpT), aquaporin Z (AqpZ) and peptide transporter (PepTso) with a C-terminal fused enhanced green fluorescent protein (eGFP)[26], were selected to cover a broad range in the PCA space and different level

Table 1 | List of RHPs

	MMA	OEGMA (500)	EHMA	SPMA	OEGMA (300)	DMAEMA	M_n[a]	M_n[b]	$Đ$[c]
RHP1	70	25	-	5	-	-	18.4	20.7	1.4
RHP2	65	25	5	5	-	-	17.9	18.3	1.5
RHP3	60	25	10	5	-	-	18.5	18.6	1.5
RHP4	50	25	20	5	-	-	36.3	27.3	1.3
RHP5	40	25	30	5	-	-	26.5	22	1.3
RHP6	20	25	50	5	-	-	25.6	20.8	1.3
RHP7	-	25	70	5	-	-	31.0	24.2	1.3
DHP1	50	25	20	5	-	-	17.3	17.9	1.1
DHP2	50	25	20	5	-	-	30.5	25.8	1.4
RHP8	60	-	10	-	15	15	15.5	10.8	1.2
RHP9	50	-	20	-	15	15	15.5	11.8	1.2
RHP10	20	-	50	-	15	15	16.4	14.6	1.1
RHP11	-	-	-	-	10	90	28.7	30.1	1.3
RHP12	50	25	20	-	-	5	22.3	27.5	1.7
RHP13	50	-	20	-	25	5	17.3	11.0	1.2
RHP14	70	-	-	-	25	5	16.0	11.1	1.2

[a] Estimated by ^{1}H-NMR and reported in kDa.
[b] Estimated by gel permeation chromatography using DMF as the mobile phase and reported in kDa.
[c] Dispersity.

of folding success without RHPs during cell-free synthesis (Fig. 2a). Experimentally, the RHP solution concentration was set to 0.2 wt%, well below the RHP's critical overlap concentration (>10 wt% for RHPs studied) to minimize crowding effects. The RHP–protein collision rate is roughly 10^5 s^{-1} so that the transient RHP–protein interactions are not diffusion-limited on the basis of the translation rate of 10–20 amino acids per second in *Escherichia coli*[27,28] and estimated RHP diffusion rate of roughly 50 μm^2 s^{-1}. In the presence of RHP1–7, a nearly twofold increase in the AqpZ-eGFP protein yield was observed (Fig. 2b) using ^{35}S-methionine labelling, and there were improvements in the AqpZ-eGFP's folding status (Fig. 2c,d and Supplementary Fig. 12). Both results are positive indications of RHPs' protein-like behaviour. RHP4 has the highest level of overlapping PCA space with proteins and was the best performer. Similar trends were observed for PepTso with RHP5-6 being the best two performers. OmpT has a fairly good folding status without RHPs (Supplementary Fig. 13); this was attributed to its lower apparent hydrophobicity compared to the other membrane proteins, as seen by its higher PC1 value. RHP1–3 with higher PC1 values were the most effective in mediating OmpT folding, whereas RHP6–7 with lower PC1 values have deleterious effects.

We further probed the interplay between RHP ensembles and biological fluids on thermal denaturation as an accelerated case of how proteins under stress sample different conformations and interact with surrounding molecules. Studies were carried out using FBS, a complex mixture of more than 1,000 proteins and other components[25]. Without RHPs, precipitates formed in the FBS solution within several days when stored at room temperature. When RHPs were added, visual inspection showed a substantial reduction in the formation of precipitates over 1 month without refrigeration. The precipitation was more obvious when the FBS was heated without RHPs for 2.5 h at 52 °C (Fig. 2e), indicative of increased protein denaturation. When different RHP ensembles were compared, RHP4 performed better than RHP7 and RHP2 in stabilizing FBS and was used for subsequent studies (Extended Data Fig. 2). When the thermally treated FBS was used as a growth supplement for in vitro cell culture of NIH3T3 fibroblast cells, there was a 19.1 ± 8.5% (mean ± s.d., n = 8) reduction in the cell viability. When 0.5 mg ml^{-1} of RHP4 ensemble was added into FBS before heating, the cell viability was retained at 93.0 ± 12.2% of control FBS without thermal treatment. The cell viability increased to 95.9 ± 10.8 and 104.8 ± 7.8% at RHP4 concentrations of 1 and 2.5 mg ml^{-1}, respectively. FBS is a biological fluid commonly used in cell culture, so these studies indicate that RHP ensembles with matching PCA spaces can interact with proteins as if they were proteins themselves. In both studies, we minimized specific factors such as molecular chaperones, osmolytes and crowding effects for protein stabilization[29–31]. These results further confirmed the ability of RHP ensembles to navigate through compositional uncertainties during random encounters.

RHP segment availability depends on local RHP–protein interactions

Each RHP ensemble includes several segments that span a range of segmental hydrophobicity and sequential arrangements within an RHP chain. Ultimately, the availability of these segments during random encounters governs the apparent protein–RHP interactions. Thus, we probe whether the spatial arrangement of segments can be influenced by their immediate environment and the segmental sequence. Solution small-angle X-ray scattering (SAXS) studies showed that RHP4 in water formed a single-chain nanoparticle, 8.8 nm in average diameter, to bury hydrophobic monomers (Fig. 3a). Proton nuclear magnetic resonance (^{1}H-NMR) studies of RHP4 in D_2O showed that the surface-exposed segments are more hydrophilic and mobile; the buried segments are more hydrophobic and have a low tendency to snorkel to the surface of globular RHP chains (Extended Data Fig. 3a), consistent with previous atomistic molecular dynamic simulations[32]. However, from ^{1}H-NMR, the full-width at half-maximum (FWHM) accounting for the end-methyl protons of the EHMA side-chain and backbone methyl protons transitioned from being fairly broad to sharp on heating from 25 to 70 °C (Fig. 3b and Extended Data Fig. 3b). In the same temperature range, the FWHM of protons from the end methyl group in the OEGMA side chains remained sharp consistently (Extended Data Fig. 3c). This result indicates buried hydrophobic residues or segments become more flexible and solvated on heating[33]. Adding dimethyl sulfoxide (DMSO) into

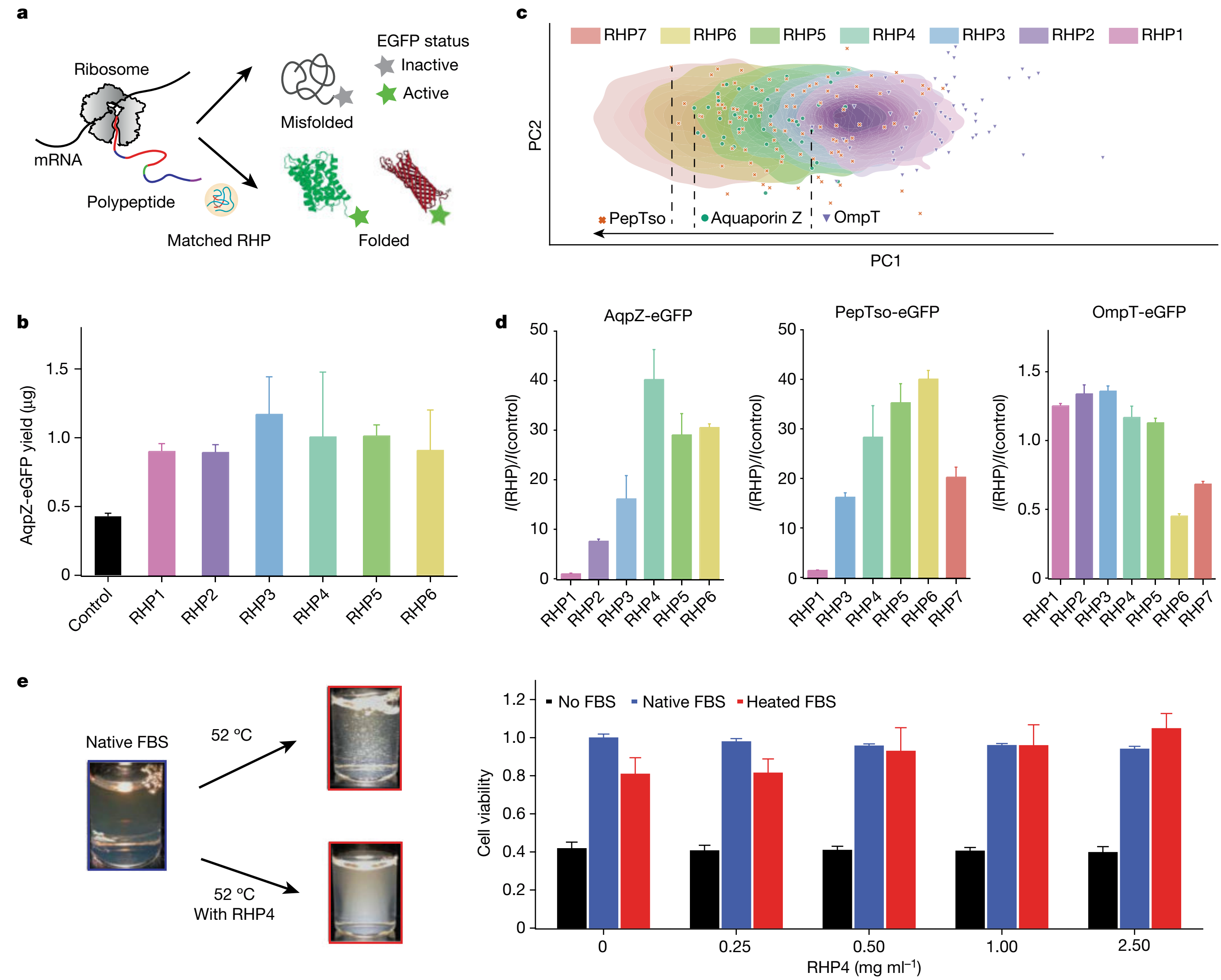


Fig. 2 | RHP/protein PCA space overlap determines their interplay. a, Scheme showing RHPs facilitate a top-down conformational sampling process of membrane proteins during protein folding. **b**, Protein yield of AqpZ-eGFP in 8 µl of cell-free reaction in the presence of 0.2 wt% RHPs ($n = 2$). Error bar is 1 s.d. **c**, PCA map showing the segment distributions for both RHP libraries and tested proteins (OmpT (β-barrel), AqpZ (α-helical) and PepTso (α-helical)). **d**, Folding status of AqpZ-eGFP, PepTso-eGFP and OmpT-eGFP in the presence of 0.2 wt% RHPs based on the eGFP fluorescence ($n = 3$). I(RHP) and I(control) are the fluorescence intensity of tested proteins in the presence and absence of RHP, respectively. Error bar is 1 s.d. **e**, The left shows the presence of RHP4 prevents FBS precipitation on heating. The right shows the cell viability of NIH3T3 as a function of RHP4 concentrations ($n = 8$; the final RHP concentrations in the cell culture media were diluted fivefold from labelled concentrations in the x axis). Error bar is 1 s.d. The cell viability was indirectly measured by metabolic viability-based assays using tetrazolium salts, MTT. At any given concentration of RHP4, RHP4 alone (no FBS), fresh FBS–RHP4 mixture (native FBS) or heated FBS–RHP4 mixture (heated FBS) was fed into the cell culture media.

the D_2O has a similar effect to heating and also causes the proton peak of the EHMA side chains to sharpen (Extended Data Fig. 4).

During random encounters, the molecules in contact with an RHP can locally solvate and lower the energy barrier to surface-expose amphiphilic or hydrophobic RHP segments[34]. No method exists to directly image the RHPs' conformations when they transiently interact with proteins, so an atomistic molecular dynamics simulation was performed. In the simulation, a non-polar hexane nanodroplet was placed near an RHP4 in water. Our simulations showed that an RHP4 globule can be fused and subsequently unravelled at the interface very quickly (roughly 100 ns) (Fig. 3c). This is by stark contrast to frozen RHP4 backbones observed in pure water[32]. There is a substantial conformational rearrangement of RHP at the hexane nanodroplet/water interface, shielding non-polar groups from exposure to water (Supplementary Figs. 14–16). When all studies are considered together, it is reasonable to conclude that an RHP chain can effectively modulate its side-chain distribution on the basis of the proteins it contacts, and can thus provide matching segments on-demand to assist proteins as they traverse back to their native states.

An RHP chain can be viewed as an equivalent freely jointed chain with segments spanning a range of segmental hydrophobicity (Fig. 3d). The exchange dynamics of each segment depend on its hydrophobicity and length, analogous to the single-chain desorption or exchange kinetics of amphiphilic block copolymers[35]. To understand how the segment length affects RHPs' conformational plasticity, we synthesized two DHPs, p(MMA-co-EHMA)-b-p(OEGMA-co-SPMA), with the same composition as RHP4, but different molecular weights (M_n (DHP1) of 17.3 kDa and M_n (DHP2) of 30.5 kDa). DHPs have multimodal distributions in the PCA space that account for 50-mers from the p(MMA-co-EHMA) block, in-between block or p(OEGMA-co-SPMA) block, respectively (Extended Data Fig. 5). The occupied PCA space shifts along both PC1 and PC2

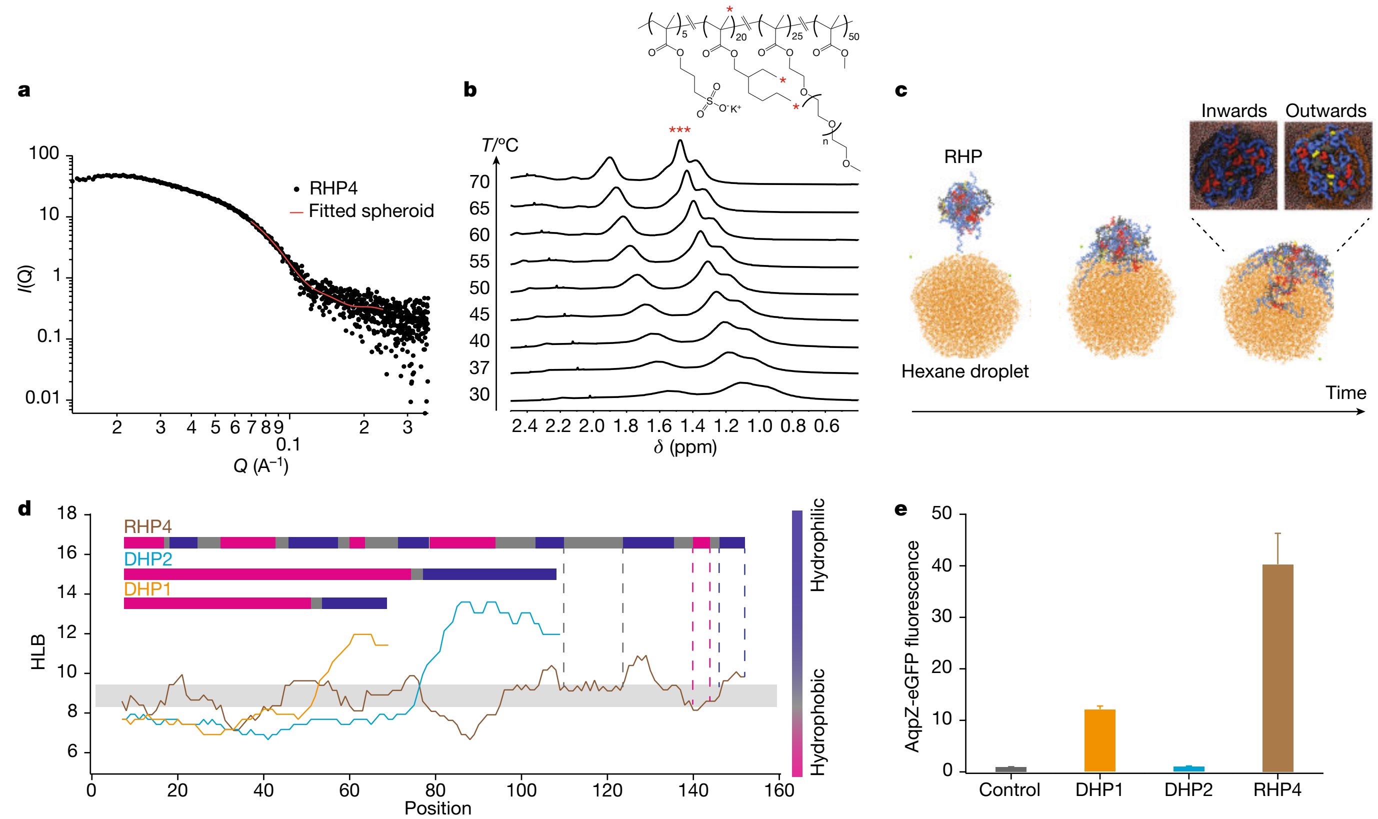


Fig. 3 | RHPs provide diverse segments with a defined range of segmental hydrophobicity to modulate transient intermolecular interactions. a, The solution SAXS profile of RHP4 in water at 5 mg ml^{-1} fitted using a spherical model with a diameter of 8.8 nm. **b**, Part of 1H-NMR spectra of RHP4 as a function of temperature (30–70 °C). **c**, RHP4 adjusts local segmental conformation when exposed to a droplet of hexane, suggesting RHP segments could be available based on the surface of protein in contact. The RHP side-chain configurations at the interface towards the hexane phase ('inwards') and towards the water phase ('outwards') are shown. The system was equilibrated over 100 ns and explicit solvent and counterions are omitted for clarity. Hexane is shown in orange with MMA, OEGMA, EHMA and SPMA in grey, blue, red and yellow, respectively. **d**, Sliding window analysis showing the segmental hydrophobicity along a chain of RHP4, DHP1 and DHP2, respectively. The hydrophobicity of each monomer along a polymer chain was evaluated by the average HLB value of a sliding window. Differences in the diversity of segments are seen between RHPs and diblock copolymers. The block length in copolymers also defines the range of segmental hydrophobicity. **e**, The folding status of eGFP-fused AqpZ in the presence of 0.2 wt% DHP1, DHP2 and RHP4, respectively ($n = 3$). The corresponding fluorescence intensity is normalized to that measured from the control experiment (no RHP). Error bar is 1 s.d.

axes, leading to a substantial deviation from the protein PCA space. In an assay of AqpZ folding status, DHP1 could increase the eGFP fluorescence intensity during translation, but did not perform as well as its RHP counterparts (Fig. 3e). Negligible enhancement of eGFP fluorescence was detected when DHP2 was used. Dynamic light scattering (DLS) measurements showed that both DHPs have bimodal size distributions with a mean diameter of 56.2 nm for DHP1 (Supplementary Fig. 17) and 145.5 nm for DHP2 (Supplementary Fig. 18), much larger than the 9.1 nm diameter of their RHP counterpart (Supplementary Fig. 19). Thus, the PC2 value is important for predicting the RHPs' tendency to form large assemblies that will compromise the PC1 similarity between RHPs and proteins. The prevalence of short hydrophobic/amphiphilic blocks in RHPs is critical to provide its conformational flexibility, thus ensuring segments' availability.

Designing new RHP ensembles as cytosol mimics

Biomacromolecules need to function coherently regardless of their own specialties. We propose that the sequence space of proteins encodes how protein mixtures behave beyond individual interactions. Therefore, a population-based design approach for synthetic protein analogues is more meaningful than pursuing molecularly precise formulations. In comparison to the model blends used at present to replicate biological multi-scale phase behaviours, RHP solutions are more relevant because they access key characteristics such as chemical diversity, compositional complexity and uncertainty[10,36–38]. As a test case, a new RHP library was synthesized to mimic cytosols with common functions including: (1) tunable microscopic LLPS under biologically relevant conditions; (2) the ability to interface and modulate protein folding and stability and (3) compartmentalization of biomolecules such as DNAs to modulate their bioavailability. Designing RHPs to fulfil many functions, each with a high level of biological importance, will test RHPs' similarities to natural proteins and the robustness of population-based design.

RHP8–14 (Table 1) were synthesized using 2–4 of the available monomers: MMA, EHMA, OEGMA (M_n = 300 Da), OEGMA (M_n = 500 Da) and DMAEMA. Two OEGMA monomers with different side-chain lengths and DMAEMA were chosen to implement and modulate temperature-dependent intermolecular interactions[39] to access LLPS under biologically relevant conditions. The RHP monomer ratios were selected to have different levels of overlap in the 2D PCA space with known proteins undergoing LLPS behaviours (from the LLPSDB database[40]). Sequence analysis showed that RHP8–10 overlap with the 2D PCA space of proteins undergoing LLPS, and RHP11 lies outside the protein space (Fig. 4a).

RHP8–10 formed spherical droplets suspended in the buffer at 47 °C for less than 1 min (Fig. 4a,b and Extended Data Fig. 6). The cloud point temperatures of RHP8, 9 and 10 solutions at 1 mg ml^{-1}

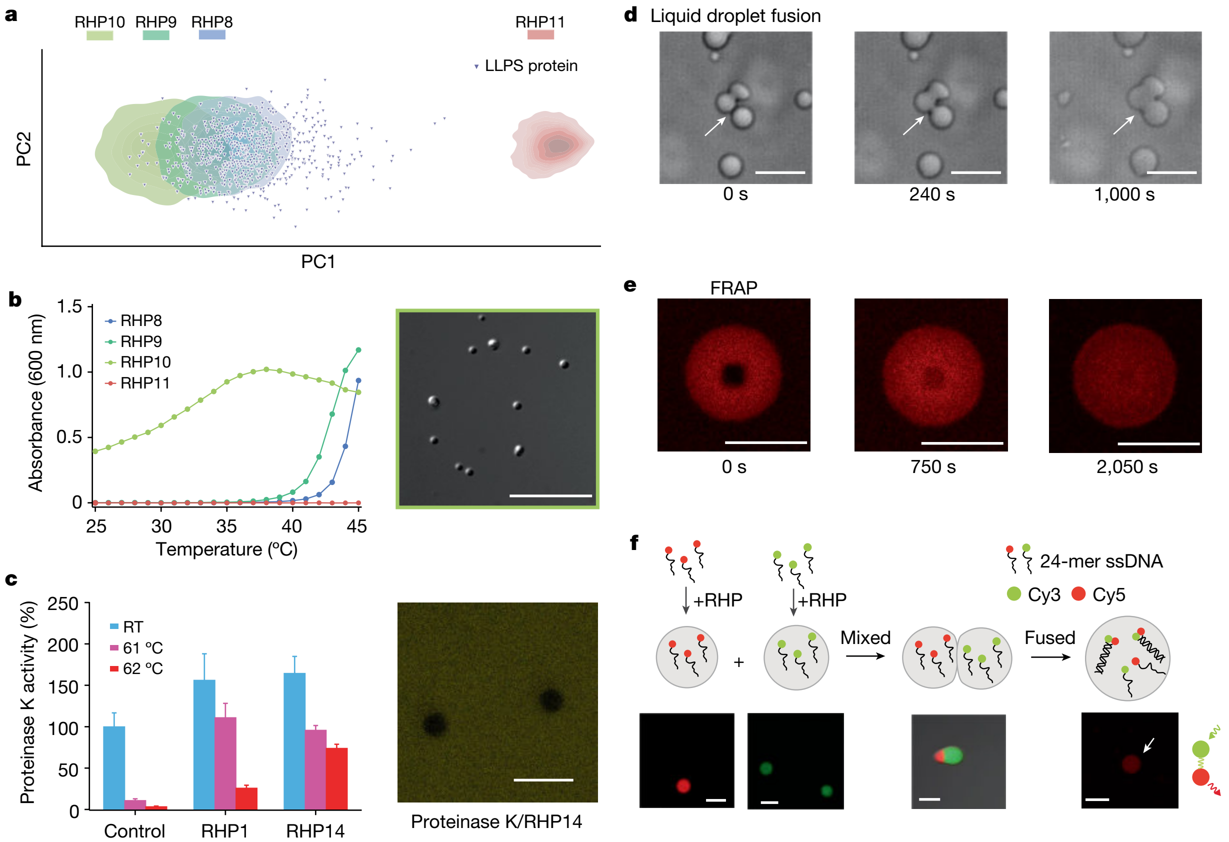


Fig. 4 | Designing RHP ensembles as synthetic mimics of cytosol. a, The PCA map of segment distributions for both RHP8–11 ensembles and proteins. **b**, The left shows the temperature-dependent turbidity for RHP8–11 ensembles (1 mg ml^{-1} in sodium phosphate buffer (50 mM, pH 7.0)). The right shows a differential interference contrast image that RHP10 ensemble forms liquid droplets at 47 °C. **c**, The left shows the normalized hydrolytic activity of ProK before and after heat treatment in the absence or presence of RHPs (n = 3). Error bar is 1 s.d. The original activity of ProK before heat treatment was set to 100%. The right shows the confocal image that the fluorescein-labelled ProK was excluded from RHP14 droplets. RT, room temperature. **d**, Spherical RHP10 droplets fuse into a larger sphere. **e**, FRAP showing the dynamic reorganization of Cy3 dye within an RHP10 droplet. **f**, The confocal images of RHP10 coacervates incubated with 3′-Cy3 ssDNA1 (24 nt; (CAGT)6), 5′-Cy5 ssDNA2 (24 nt; (ACTG)6) or ssDNA1–ssDNA2 duplex. Scale bars, 5 µm.

(sodium phosphate buffer, 50 mM, pH 7.0) were determined to be 43, 40 and <25 °C, respectively. Under the same buffer condition, RHP11 showed no phase separation even up to 70 °C. Thus, matching the PCA space between proteins and RHP ensembles is an effective approach to replicate the collective behaviours of proteins, such as LLPS.

We studied RHP–protein interactions with and without LLPS using RHP12 and RHP13. Both have the same monomer ratio as RHP4, except that SPMA was replaced by DMAEMA and the side-chain length in OEGMA was varied. RHP12 has a cloud point temperature of roughly 67 °C and the cell-free AqpZ-eGFP folding assay at 37 °C confirmed that RHP12 (0.2 wt%) can facilitate AqpZ folding. In comparison, RHP13, which has a cloud point temperature lower than 25 °C, underwent LLPS at 37 °C and showed no effect in the AqpZ-eGFP folding (Extended Data Fig. 7). A control experiment showed that RHP13 did not interfere with eGFP expression or folding post translation. We also assessed how the formation of liquid droplets might affect proteins on thermal denaturation using a common enzyme, proteinase K (ProK). RHP1 has the most PCA space overlap with ProK (Supplementary Fig. 20) and was the best performer as ProK's thermal protectant. RHP1's monomer ratio was adopted to synthesize RHP14 by replacing SPMA with DMAEMA to induce LLPS. In the presence of RHP14, ProK retained 74 ± 5% enzyme activity after being heated at 62 °C for 10 min. In the control experiments without RHPs, ProK lost roughly 97% of its enzymatic activity (Fig. 4c). RHP14 has a cloud point temperature of 33 °C (Extended Data Fig. 8). Confocal studies showed that the fluorescein-labelled ProK was excluded from RHP14 liquid droplets (Fig. 4c). In comparison, only 25% of ProK activity was retained with RHP1, which showed no LLPS. Together, these results indicated that once the LLPS occurred, the hydrophobic RHP segments and some amphiphilic RHP segments became inaccessible to proteins outside the droplets. These segments are the key to mediating membrane protein–water interactions and assisting membrane protein folding. Their absence reduced the probability of ProK misfolding during thermal denaturation. These results also indicated these hydrophobic or amphiphilic segments provided the driving forces to form the liquid droplets. NMR studies showed no difference in the monomer compositions between the RHPs inside of the liquid droplets and the original RHP ensemble. Thus, it is reasonable to speculate that the interchain interactions within the droplets are dominated by the apparent segmental hydrophobicity within the RHP chains instead of the whole chain.

RHP10 formed many small droplets that ranged from hundreds of nanometres to a few micrometres in size. They diffused, collided with one another and fused together within a few minutes (Fig. 4d). Cyanine 3 amine (Cy3) dye was concentrated in the droplets and used to probe the local viscosity within droplets through fluorescence recovery after photobleaching (FRAP). Nearly complete fluorescence recovery was observed with a characteristic recovery time of 385 ± 15 s (Fig. 4e and Extended Data Fig. 9). This time scale is comparable to that of MEG-3 proteins (128–384 s) in P granules[41,42]. The RHP droplets have a fluidity similar to membraneless organelles, and offer a promising path towards their synthetic mimics.

We propose that the design rules used here to modulate RHP–protein interactions can also be applied to tailor how RHPs interact with DNA. Fundamentally, RHPs enable us to evaluate energetic competition between a wide range of interactions, going beyond the solely electrostatic interactions commonly studied in coacervation. When 1 μM of 24-mer single-stranded DNA (ssDNA, 24 nt) was added to the RHP10 solution (1 mg ml^{-1}, sodium phosphate buffer, 50 mM, pH 7.0), ssDNA was selectively encapsulated inside the liquid droplets at 37 °C (Fig. 4f). Similar results were observed for a 24-base pair fluorescence resonance energy transfer (FRET)-pair labelled double-stranded DNA (dsDNA). FRET was observed inside the liquid droplets, indicating the absence of dsDNA dissociation. When liquid droplets carrying complementary FRET-pair labelled ssDNA were mixed, these droplets fused. Subsequently, complementary ssDNA strands formed duplexes and emitted FRET over the course of 25 min. These results confirmed that the RHP–DNA interactions are strong enough for selective partition but do not interfere with specific DNA pairing interactions. Thus, designed RHPs can capture the range of intra- and intermolecular interactions required to be compatible with many biological processes occurring inside of cytosol, making them viable building blocks towards bio- or abiotic materials.

The present studies clearly demonstrate the feasibility of designing heteropolymers as an ensemble to mirror protein mixtures in biological fluids despite the inexact formulations of these complex blends. The segmental sequence information of proteins can guide statistic sequence control in RHPs to holistically replicate protein functions. Being orthogonal to molecular-precision-driven design, the population-based approach is actually advantageous to navigate through chemical diversity and unpredictable fluctuations abundant in protein's native environments and, ultimately, to interface synthetic materials and biological systems seamlessly.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-05675-0.

1. Fulton, A. B. How crowded is the cytoplasm? *Cell* **30**, 345–347 (1982).
2. Lewandowski, J. R., Halse, M. E., Blackledge, M. & Emsley, L. Direct observation of hierarchical protein dynamics. *Science* **348**, 578–581 (2015).
3. Dill, K. A. Dominant forces in protein folding. *Biochemistry* **29**, 7133–7155 (1990).
4. Wirth, A. J., Platkov, M. & Gruebele, M. Temporal variation of a protein folding energy landscape in the cell. *J. Am. Chem. Soc.* **135**, 19215–19221 (2013).
5. Keskin, O., Gursoy, A., Ma, B. & Nussinov, R. Principles of protein–protein interactions: what are the preferred ways for proteins to interact? *Chem. Rev.* **108**, 1225–1244 (2008).
6. Golumbfskie, A. J., Pande, V. S. & Chakraborty, A. K. Simulation of biomimetic recognition between polymers and surfaces. *Proc. Natl Acad. Sci. USA* **96**, 11707–11712 (1999).
7. Minton, A. P. How can biochemical reactions within cells differ from those in test tubes? *J. Cell Sci.* **119**, 2863–2869 (2006).
8. DePristo, M. A., Weinreich, D. M. & Hartl, D. L. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat. Rev. Genet.* **6**, 678–687 (2005).
9. Rutherford, S. L. Between genotype and phenotype: protein chaperones and evolvability. *Nat. Rev. Genet.* **4**, 263–274 (2003).
10. Alberti, S., Gladfelter, A. & Mittag, T. Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates. *Cell* **176**, 419–434 (2019).
11. Tamasi, M. J. et al. Machine learning on a robotic platform for the design of polymer-protein hybrids. *Adv. Mater.* **34**, e2201809 (2022).
12. Webb, M. A., Jackson, N. E., Gil, P. S. & de Pablo, J. J. Targeted sequence design within the coarse-grained polymer genome. *Sci. Adv.* **6**, eabc6216 (2020).
13. Panganiban, B. et al. Random heteropolymers preserve protein function in foreign environments. *Science* **359**, 1239–1243 (2018).
14. Jiang, T. et al. Single-chain heteropolymers transport protons selectively and rapidly. *Nature* **577**, 216–220 (2020).
15. Nguyen, T. D., Qiao, B. F. & de la Cruz, M. O. Efficient encapsulation of proteins with random copolymers. *Proc. Natl Acad. Sci. USA* **115**, 6578–6583 (2018).
16. Bateman, A. et al. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
17. Monera, O. D., Sereda, T. J., Zhou, N. E., Kay, C. M. & Hodges, R. S. Relationship of sidechain hydrophobicity and alpha-helical propensity on the stability of the single-stranded amphipathic alpha-helix. *J. Pept. Sci.* **1**, 319–329 (1995).
18. Kramer, M. A. Nonlinear principal component analysis using autoassociative neural networks. *AlChE J.* **37**, 233–243 (1991).
19. Vidal, R., Ma, Y. & Sastry, S. S. *Generalized Principal Component Analysis* 25–62 (Springer, 2016).
20. Smith, A. A. A., Hall, A., Wu, V. & Xu, T. Practical prediction of heteropolymer composition and drift. *ACS Macro Lett.* **8**, 36–40 (2019).
21. Bustamante, C. J., Chemla, Y. R., Liu, S. & Wang, M. D. Optical tweezers in single-molecule biophysics. *Nat. Rev. Methods Primers* **1**, 25 (2021).
22. Ritchie, D. B. & Woodside, M. T. Probing the structural dynamics of proteins and nucleic acids with optical tweezers. *Curr. Opin. Struct. Biol.* **34**, 43–51 (2015).
23. Suzuki, Y. & Dudko, O. K. Single molecules in an extension clamp: extracting rates and activation barriers. *Phys. Rev. Lett.* **110**, 158105 (2013).
24. Zhao, Q. Nature of protein dynamics and thermodynamics. *Rev. Theor. Sci.* **1**, 83–101 (2013).
25. Gstraunthaler, G., Lindl, T. & van der Valk, J. A plea to reduce or replace fetal bovine serum in cell culture media. *Cytotechnology* **65**, 791–793 (2013).
26. Drew, D., Lerch, M., Kunji, E., Slotboom, D. J. & de Gier, J. W. Optimization of membrane protein overexpression and purification using GFP fusions. *Nat. Methods* **3**, 303–313 (2006).
27. Shastry, M. C. R. & Roder, H. Evidence for barrier-limited protein folding kinetics on the microsecond time scale. *Nat. Struct. Biol.* **5**, 385–392 (1998).
28. Milo, R. & Phillips, R. *Cell Biology by the Numbers* (Garland Science, Taylor & Francis Group, 2016).
29. Saio, T., Guan, X., Rossi, P., Economou, A. & Kalodimos, C. G. Structural basis for protein antiaggregation activity of the trigger factor chaperone. *Science* **344**, 597–610 (2014).
30. Zhou, H. X., Rivas, G. N. & Minton, A. P. Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences. *Annu. Rev. Biophys.* **37**, 375–397 (2008).
31. Street, T. O., Bolen, D. W. & Rose, G. D. A molecular mechanism for osmolyte-induced protein stability. *Proc. Natl Acad. Sci. USA* **103**, 13997–14002 (2006).
32. Hilburg, S. L., Ruan, Z. Y., Xu, T. & Alexander-Katz, A. Behavior of protein-inspired synthetic random heteropolymers. *Macromolecules* **53**, 9187–9199 (2020).
33. Terashima, T., Sugita, T., Fukae, K. & Sawamoto, M. Synthesis and single-chain folding of amphiphilic random copolymers in water. *Macromolecules* **47**, 589–600 (2014).
34. Shu, J. Y. et al. Amphiphilic peptide-polymer conjugates based on the coiled-coil helix bundle. *Biomacromolecules* **11**, 1443–1452 (2010).
35. Choi, S. H., Lodge, T. P. & Bates, F. S. Mechanism of molecular exchange in diblock copolymer micelles: hypersensitivity to core chain length. *Phys. Rev. Lett.* **104**, 047802 (2010).
36. Quiroz, F. G. & Chilkoti, A. Sequence heuristics to encode phase behaviour in intrinsically disordered protein polymers. *Nat. Mater.* **14**, 1164–1171 (2015).
37. Hyman, A. A., Weber, C. A. & Juelicher, F. Liquid-liquid phase separation in biology. *Annu. Rev. Cell Dev. Biol.* **30**, 39–58 (2014).
38. Sing, C. E. & Perry, S. L. Recent progress in the science of complex coacervation. *Soft Matter* **16**, 2885–2914 (2020).
39. Fournier, D., Hoogenboom, R., Thijs, H. M. L., Paulus, R. M. & Schubert, U. S. Tunable pH- and temperature-sensitive copolymer libraries by reversible addition-fragmentation chain transfer copolymerizations of methacrylates. *Macromolecules* **40**, 915–920 (2007).
40. Wang, X. et al. LLPSDB v2.0: an updated database of proteins undergoing liquid-liquid phase separation in vitro. *Bioinformatics* **38**, 2010–2014 (2022).
41. Putnam, A., Cassani, M., Smith, J. & Seydoux, G. A gel phase promotes condensation of liquid P granules in *Caenorhabditis elegans* embryos. *Nat. Struct. Mol. Biol.* **26**, 220–226 (2019).
42. McSwiggen, D. T., Mir, M., Darzacq, X. & Tjian, R. Evaluating phase separation in live cells: diagnosis, caveats, and functional consequences. *Gene Dev.* **33**, 1619–1634 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2023

Methods

Materials

Chemicals were purchased from Sigma Aldrich Chemical Co. or Fisher Scientific International, Inc. unless otherwise noted. Water was purified by a Milli-Q water filtration station (18.2 ΩM cm^{-1}) before use. Azobisisobutyronitrile (AIBN) was recrystallized from ethanol before use. Inhibitors were removed using cryodistillation (MMA, EHMA) or by passing through a short column of neutral alumina (ethylene glycol methyl ether methacrylate) (M_n roughly 500 g mol^{-1}). 3-SPMA (98%) before polymerization. Ethyl-2(phenylcarbanothioylthio)-2-phenylacetate, trioxane (internal standard for 1H-NMR analysis) and dimethylformamide (DMF, solvent) were used without further purification. Cell-free transcription/translation system (PURExpress) was purchased from New England BioLabs Inc. Plasmid Aqpz-GFP was a gift from D.L. Minor, Jr. (University of California, San Francisco). Plasmid pWaldoGFPe_PepTSo was a gift from S. Iwata and S. Newstead (Addgene plasmid no. 58334). pET28-OmpT was a gift from N. Kelleher (Addgene plasmid no. 68862). Anti-GFP antibody (B2) was purchased from Santa Cruz Biotechnology. Goat antimouse alkaline phosphatase secondary antibody conjugate was purchased from Bio-Rad Laboratories. End-modified oligonucleotides were purchased from IDT Co. ProK was purified by a desalting column before use. Dulbecco's Modified Eagle Medium (DMEM) and cell culture plates were purchased from Corning Inc. The FBS and phosphate buffered saline were purchased from Gibco Inc. The 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) and pancreatin were purchased from Invitrogen Co. The NIH3T3 cell line was provided by the UCB Cell Culture Facility.

RHP synthesis

The synthesis of RHPs was carried out using reversible addition-fragmentation chain-transfer polymerization as in our previous work. Polymerization solutions were prepared by mixing the requisite amounts of reagents and added into 25 ml glass Schlenk. The solutions were subject to three freeze–pump–thaw cycles before the Schlenk was placed into the oven. The polymerization solutions were held at 70 °C, typically for 4–8 h. Global monomer conversion was determined by ^{1}H-NMR on crude reaction mixtures in DMSO-d_6 with trioxane as an internal standard. The polymerization solution was then precipitated by dropwise addition to rapidly stirred pentane. The resultant polymer was then dissolved in water and transferred to a 2,000 molecular weight cut-off (MWCO) dialysis bag and dialysed against distilled water for 3 days. The purified polymer was then subject to lyophilization.

Diblock copolymer *p*(MMA-co-EHMA)-b-*p*(OEGMA-co-SPMA) synthesis

An example of the preparation of *p*(MMA-r-EHMA)$_{53}$-b-*p*(OEGMA-r-SPMA)$_{22}$ (M_n (NMR) = 17,327 g mol^{-1}) is given.

***p*(MMA-r-EHMA) macro-CTA synthesis.** MMA (1,072.71 mg, 10.71 mmol), 2-EHMA (849.86 mg, 4.29 mmol), CTA (2-cyano-2-propyl-4-cyanobenzodithioate, 32.13 mg, 0.13 mmol), AIBN (2.14 mg, 0.013 mmol) and DMF (1 ml) were mixed in a 20 ml glass vial. The reaction mixture was deoxygenated by N_2 flow for 10 min, placed in an oven and allowed to react at 70 °C, typically for 3 h. After polymerization, the crude product was reprecipitated in pentane three times, redissolved in tetrahydrofuran (THF) and transferred to a clean vial. The solvent was removed by evaporation under reduced pressure. The purified polymer was then dried in vacuo overnight.

***p*(MMA-co-EHMA)-b-*p*(PEGMA-co-SPMA) synthesis.** OEGMA (1,027.41 mg, 2.05 mmol), 3-SPMA (101.29 mg, 0.41 mmol), *p*(MMA-r-EHMA)$_{53}$ macroRAFT agent (335.743 mg, 0.0476 mmol, M_n = 7,056 g mol^{-1}), AIBN (0.781 mg, 0.00476 mmol) and DMF (1 ml) were added and mixed in a 20 ml glass vial. The reaction mixture was deoxygenated by N_2 flow for 10 min, placed in an oven and allowed to react at 70 °C for roughly 3 h. After polymerization, the crude product was reprecipitated in pentane three times, redissolved in THF and dialysed against a 12:1 mixture of THF and water for 3 days. The solvent was removed by evaporation under reduced pressure. The purified polymer was then subject to lyophilization.

Polymer characterization

Molecular weight distribution curves, number-average (M_n) and weight-average (M_w) molar mass and dispersity ($Đ = M_w/M_n$) of copolymers were measured by gel permeation chromatography using an Agilent 1260 Infinity series instrument outfitted with two Agilent PolyPore columns (300 × 7.5 mm). DMF with 0.05 M LiBr was used as the eluent at 0.7 ml min^{-1} at 50 °C. The columns were calibrated against Poly (ethylene glycol) standards or by light scattering. Analyte samples at 2 mg ml^{-1} were filtered through 0.2 µm polytetrafluoroethylene membranes (VWR) before injection (20 µl). ^{1}H-NMR spectra were recorded in DMSO-d_6 and D_2O on a Bruker Avance 400 spectrometer (400 MHz) using a 5 mm *Z* gradient BBO probe or on a Bruker Avance AV 600 spectrometer (600 MHz) using a *Z* gradient Triple Broad Band Inverse detection probe.

Cell-free protein synthesis

Cell-free protein synthesis was carried out according to PURExpress manual with modifications. In 25 µl of reaction, RHP of the desired amount was added to the ribosome solution and incubated on ice for 30 min. The polymer/ribosome samples were mixed with solution A and B containing all other required components including enzymes, RNAs, energy and nutrient molecules. Next, 20 units of RNase inhibitor and 200 ng plasmids for AqpZ-eGFP, PepTso-eGFP and OmpT-eGFP were added, and the mixtures were incubated at 37 °C for 4 h to complete membrane protein synthesis. The resulting products were stored at −20 °C.

Kinetics of protein synthesis and western blot analysis

Kinetics of cell-free synthesis were monitored by measuring the fluorescence intensity of the C-terminal GFP tag on a Tecan I-control infinite 200 plate reader. Cell-free protein synthesis mixtures were incubated in a 384-well plate at 37 °C, and the GFP fluorescence (excitation/emission 488 nm/520 nm) was recorded over 4 h. For the western blot, proteins in the cell-free protein synthesis samples were first separated by SDS–PAGE. Protein mass standards were used (Spectra multicolour broad range protein ladder, Thermo Fisher). The proteins were then transferred to a polyvinylidene difluoride membrane. Multicolour protein mass standards were visible on the membrane after successful transfer. The membrane was then blocked with 3% BSA, incubated with 1:2,000 mouse anti-GFP antibody and washed. Following incubation with 1:4,000 alkaline phosphatase conjugated goat antimouse secondary antibody, band colours were developed by using 5-bromo-4-chloro-3-indolyl phosphate and nitroblue tetrazolium.

Protein yield determination

The protein yield of cell-free synthesis was determined by measuring the ^{35}S-Methionine incorporated samples in a scintillation counter. EasyTag^{35}S-Methionine (7.2 pmol) was added in the 24 µl of cell-free synthesis solution, and the mixtures were incubated at 37 °C for 4 h to incorporate ^{35}S-Methionine into the proteins. Then 8 µl of the labelled cell-free reaction was mixed with 100 µl of 1 M NaOH and incubated at room temperature for 10 min. Next, 800 µl cold TCA/CAA mix (25% trichloroacetic acid/2% casamino acids) was added to the sample. The sample was briefly vortexed, then incubated on ice for 5 min. The precipitated proteins were collected on

a paper filter by vacuum filtration and measured in a scintillation counter.

Hydrolytic activity assay for ProK

The hydrolytic activity of ProK with different RHPs was measured by monitoring the absorbance at 410 nm to detect *p*-nitroaniline, which is released after the hydrolysis of 4-nitrophenyl butyrate. For the thermal denaturation assay, a mixture solution containing 0.07 mg ml^{-1} ProK and as required, 2 mg ml^{-1} RHP in sodium phosphate buffer (50 mM, pH 7.0) was incubated at elevated temperature for 30 min. The sample containing 10 µl of the aforementioned solution, 1 µl of 4-nitrophenyl butyrate (50 mM in methanol) and 239 µl of Tris buffer (50 mM, pH 8.0) was analysed in absorption at 410 nm.

Cell culture

The NIH3T3 cell line was received frozen. The cells were thawed, diluted in DMEM (Gibco, 4.5 g l^{-1} glucose, L-glutamine, sodium pyruvate, 10% FBS) and incubated at 37 °C and 5% CO_2. NIH3T3 cell line was divided every 3 days.

MTT assay

RHP was added to FBS at concentrations of 0.25, 0.5, 1 and 2.5 mg ml^{-1}. The mixed solution was incubated at 52 °C for 2.5 h. The supernatant was collected after centrifugation (12,000*g*, 10 min) and diluted by a factor of 5 in DMEM (Gibco, 4.5 g l^{-1} glucose, L-glutamine, sodium pyruvate, denoted SolA). NIH3T3 cells were seeded in 96-well plates at 10,000 cells per well (culture volume 100 µl per well) and incubated at 37 °C and 5% CO_2 for 16 h. The cells were then fed with SolA and incubated at 37 °C and 5% CO_2 for 24 h. The cell culture media was removed and 50 µl of MTT reagent solution (0.5 mM in DMEM, Gibco, phenol red free, 4.5 g l^{-1} glucose, L-glutamine, sodium pyruvate) was added per well. The plate was incubated at 37 °C and 5% CO_2 for 30 min. Next, 150 µl of DMSO was added per well and the plate was shaken to dissolve formazan completely. The absorbance at 570 nm was recorded using Infinite M200 microplate reader (Tecan).

Differential scanning calorimetry (DSC)

DSC measurements were implemented using the MicroCal VP-DSC system (Malvern Panalytical). All samples were prepared in 50 mM Sodium Phosphate buffer (pH 7.0) with 0.28 mg ml^{-1} ProK and as required, 1.5 mg ml^{-1} RHP. The samples were degassed for 8 min using the MicroCal ThermoVac (Malvern Panalytical) before insertion into the sample cell. When the pressure in the sample holder had stabilized at roughly 28 psi, the cells were cooled to 10 °C, followed by a 15 min wait time. The thermograms were produced by measuring the difference in heat capacity between the sample solution and the reference buffer solution as the temperature was increased from 10 to 90 °C at a rate of 1 °C min^{-1}. Following the retrieval of the thermograms, the buffer–buffer reference thermograms were subtracted using OriginLab software, as was any linear baseline observed.

DLS

DLS measurements were conducted on a Brookhaven BI-200SM Light Scattering System at a 90° scattering angle. The concentration for each measurement was 5 mg ml^{-1} of RHP or 0.5 mg ml^{-1} of DHP.

Diffusion-limited collision rate estimation (d*n*/d*t*)

$$dn/dt = j(r)4\pi r^2 = 4\pi\ \mathrm{Drc}_\infty = k_{on}c_\infty$$

where *n* is the number of molecules, *t* is time, *j* is the flux, *r* is the radius, D is the diffusion coefficient, r is the radius of the protein of interest, K_{on} is the polymer-protein association rate, and c_∞ is the bulk concentration of polymers.

As proteins are around 4 nm in diameter,

$$k_{on} = 4\pi \mathrm{Dr} = 4\pi \times 50\ \mu\mathrm{m}^2/\mathrm{s} \times 2\times 10^{-3}\ \mu\mathrm{m} \times 6.02 \times 10^{23}\ \text{molecules per mol} \times 10^{-15}\ \mathrm{l}\ \mu\mathrm{m}^{-3} \approx 10^9\ \mathrm{s}^{-1}\mathrm{M}^{-1}.$$

At 2 mg ml^{-1} of RHP4 (M_n = 36 kDa),

$$dn/dt = k_{on}c_\infty = 10^9\ \mathrm{s}^{-1}\mathrm{M}^{-1} \times 2\ \mathrm{g\ l}^{-1}/(3.6\times 10^4\ \mathrm{g\ mol}^{-1}) \approx 10^5\ \mathrm{s}^{-1}.$$

Critical overlap concentration of RHP4

$$c^* \cong 3\mathrm{M}/4\pi N_A R_g^3 \cong 17\ \mathrm{wt\%}.$$

where c^* is the critical overlap concentration, M is the average molecular weight of the polymer, R_g is the radius of gyration of the polymer, and N_A is Avogadro's number.

SAXS

SAXS was carried out at beamline 8-ID-E at the Advanced Photon Source, Argonne National Laboratory. Samples were dissolved in water at a range of concentrations from 0.2 to 2 wt%. Samples were measured in 2 mm boron-rich thin-walled capillary tubes and subject to several short exposures (5 s for each time). 2D scattering results were azimuthally averaged to produce one-dimensional SAXS profiles. Superimposable profiles were averaged and then subtracted from the background data. The radius of gyration (R_g) of an RHP was obtained from the Guinier plot by fitting the scattering to the following equation, where q is the scattering vector, *I*(q) is the scattering intensity at q, and *I*(0) is the intensity of incident beam:

$$\ln(I(\mathrm{q})) = \ln(I_0) - (R_g^2/3)q^2$$

Molecular dynamics simulation

The final frame from an equilibrated sequence (sequence 6) used in previous work[32] was extracted and solvated in a system containing a pre-equilibrated hexane droplet, four potassium counterions and 44,996 water molecules. The droplet is composed of 1,503 hexane molecules and was formed with 63,898 water molecules through 2 ns of equilibration and 40 ns of production simulation. The combined system was equilibrated for 2 ns and then run at production setpoints for 100 ns. All simulations followed the methods and parameters used in previous work, adopting the monomer parameterization methods for hexane molecules. VMD was used for visualization of the resulting trajectories[43]. Solvent accessible surface area (SASA) calculations were performed with Amber19's LCPO default parameters[44]. SASA for the hexane phase is taken as the SASA for the trajectory stripped of all non-polymer atoms (total RHP SASA) minus the SASA for the as-is trajectory (water accessible RHP SASA).

Synthesis of oligonucleotide-RHP conjugate

Buffer A (storage buffer): 20 mM sodium phosphate, pH 7.83
Buffer B: 200 mM sodium phosphate, 300 mM NaCl, pH 7.24
Buffer C: 100 mM sodium phosphate, 1,500 mM NaCl, pH 7.24
DNA1: GTCGCTCTCTCATGCAGAATCCCA, 1 mM in buffer A
DNA2: CTGCTGGGGCAAACCAGCGTGGAC, 1 mM in buffer A

Synthesis of RHP with dual end-functional groups of –SH and –N_3. An azido-modified chain-transfer agent (2-(dodecylthiocarbonothioylthio)-2-methylpropionic acid 3-azido-1-propanol ester) was used to synthesize RHP (RHP-N_3). Before conjugation, 1 µl of hydrazine was added into 200 µl of RHP solution (20.30 mg, in THF). The mixture was placed on a shaker for 0.5 h. RHP was purified by an Amicon-3K ultrafilter (3,000 MWCO) using de-ionized water six times.

Synthesis of DNA1-RHP conjugate through thiol-maleimide addition. Here, 1.5 mg of sulfo-SMCC[45] was dissolved in 100 μl of H_2O (heated up to 50 °C for clear solution), followed by the addition of 100 μl of buffer B. Then 10 μl of 3′-amine-modified DNA1 was added. The solution was placed on a shaker for 2 h at room temperature. The excess sulfo-SMCC was removed by an Amicon-3K ultrafilter (3,000 MWCO) using buffer B six times. Then, 5 nmol of purified sulfo-SMCC-DNA1 (90 μl in buffer C) was mixed with roughly 2 mg of RHP (10 μl in DMF). The reaction was placed on a shaker overnight at room temperature. After the reaction, the excess oligonucleotide was removed by Amicon-3K by washing with de-ionized water six times (Supplementary Fig. 21).

Synthesis of DNA1-RHP-DNA2 conjugate through azide-alkyne cycloaddition. Here, 10 μl of 5′-hexynyl-modified DNA2 (1 mM in buffer A) was mixed with 4 μl of DNA1-RHP conjugate (roughly 1 nmol), followed by the addition of 17 μl of DMF. Then 10 μl of fresh click-reaction solution (10 mM $CuSO_4$, 50 mM TBTA, DMF/H_2O = 1:1) and 3 μl of sodium ascorbate (100 mM in water) were added. The mixture was placed on a shaker overnight at room temperature. The excess oligonucleotide was removed by an Amicon-3K ultrafilter (3,000 MWCO) using water six times.

Single-molecular force spectroscopy

Single-molecule force-extension measurements were carried out in a dual-trap optical tweezers instrument equipped with a 1,064 nm trapping laser[46] at a trap stiffness of 0.1 to 0.2 pN nm^{-1}. First, biotinylated double-stranded (3 kilobases (kb)) DNA handles with 24 nucleotide 5′ or 3′ terminal single-stranded overhangs complementary to DNA1 or DNA2, respectively, were deposited on streptavidin-coated polystyrene beads (1 μm). Individual RHP chains were captured between trapped beads by hybridization of their conjugated DNA to the overhangs in buffer containing 20 mM Tris·HCl (pH 7.2), 100 mM KCl, 10 mM $MgCl_2$ and 5 mM beta-mercaptoethanol. Pulling and relaxing force ramps were performed at a rate of 100 nm s^{-1} from less than 1 pN to up to 25 pN and repeated several times for each molecule. Force-extension data were collected at 667 Hz. For molecules that showed a rip in their FEC, passive-mode (constant trap position) measurement was also performed. Unfolding work was calculated by integrating each FEC and subtracting the contribution of the worm-like-chain behaviour of bare 6 kb handle DNA and unfolded RHP.

RHP sequence generation

An in-house sequence simulator called Compositional Drift[20] was used to generate 15,000 chains for each RHP ensemble (DP = 50).

Training dataset

To train the sequence autoencoder, 30,000 membrane protein sequences and 30,000 globular protein sequences with 50% identity threshold were collected from the UniProt database[16]. Sequences with uncommon amino acids were discarded. Each protein was reduced into four monomer codes. The assignment of each residue to its monomer equivalent is listed in Extended Data Table 1 and Supplementary Tables 1 and 2. For each protein sequence, a set of consecutive protein motifs were collected by moving a 50-residue long window over each sequence with 15-residue long step size, resulting in a total of 1,046,845 training sequence motifs.

Latent variable model

An in-house latent variable model was developed to perform sequence dimensionality reduction and to learn protein/RHP sequence representations. The model was implemented on the basis of a typical autoencoder architecture with an extra regression module to force the latent space to learn sequence hydrophilic-lipophilic balance (HLB) distributions. Each protein sequence motif (in its RHP equivalent form) was first one-hot encoded and then passed through the encoder. The encoder embedded each sequence into a 16-dimensional latent vector, z, which was then fed into two parallel branches: the decoder and the regressor. The decoder intended to reconstruct the input sequence from z, whereas the regressor predicted the HLB values of an input sequence from z. The loss function was designed to be a weighted sum of the departure (for example, cross entropy) of the original and reconstructed sequence and the mean squared error of predicted and true HLB values. A meaningful low-dimensional latent space that captures both sequential features and HLB distributions was obtained by optimizing the reconstruction and regression loss together.

Both the encoder and the decoder were implemented with simple multilayer perceptrons. Each had three fully connected layers with 256, 128 and 64 hidden units. The regression module had two fully connected layers with 16 hidden units. Rectified linear unit non-linear functions were used throughout the network, except that Sigmoid activation was used in the output layer of the decoder. The model was trained using the ADAM optimizer with a learning rate of 0.001. A learning rate scheduler was used to reduce the learning rate when the validation loss stopped improving. All model hyperparameters were optimized with Weights and Biases[47]. We also implemented a LSTM-based autoencoder, which demonstrated similar performance as the simpler autoencoder variant model.

Coacervation

The lyophilized RHP was dissolved in Mill-Q water (30 mg ml^{-1}). Then 1 μl of RHP solution was mixed with 29 μl of sodium phosphate buffer (50 mM, pH 7.0). The coacervation was triggered by incubating the solution at 47 °C. The milky colour appeared in less than 1 min and coacervation was confirmed by bright-field microscopy. For ssDNA partitioning, 1 mg ml^{-1} RHP and 1 μM ssDNA (50 mM sodium phosphate buffer, pH 7.0) were incubated at 37 °C for 10 min before imaging. Before the dsDNA partitioning, two complementary ssDNA (5 μM for each) were mixed and heated at 95 °C for 2 min, then immediately cooled at 4 °C for 5 min. The resulting dsDNA solution were diluted to 1 μM and mixed with 1 mg ml^{-1} of RHP. The solution was incubated at 37 °C for 10 min before imaging.

The sequence of ssDNA is shown below:
/5Cy5/ACTGACTGACTGACTGACTGACTG;
CAGTCAGTCAGTCAGTCAGTCAGT/3Cy3Sp/.

Microscopy

The differential interference contrast microscopy was performed using a Zeiss AxioImager M2 microscope equipped with a Zeiss Plan-Neofluar Ph3 ×100/1.30 oil-immersion objective and QImaging Retiga 1350EX 1.4MP Monochrome CCD Microscope Camera. The DNA partitioning, RHP droplet fusion and FRAP was imaged using a Zeiss LSM 880 FCS confocal microscope with X-Cite 120LED illumination system and GaAsP photon counting detector. For ssDNA partitioning, illumination was provided by a laser with the wavelength of 514 nm (for Cy3 or Cy3-Cy5 FRET pair) or 633 nm (for Cy5). Images were averaged from eight consecutive images and were of the format 1,024 × 1,024 pixels at an 8-bit depth. RHP droplet fusion was imaged every 2 s under transmitted light in a format of 512 × 512 pixels at an 8-bit depth.

FRAP

Here, 50 μM of Cy3 amine (Lumiprobe, 410C0) was used to probe FRAP within the RHP droplets. The pinhole size was set to 1.5 μm section. FRAP measurement was performed by selecting a thin square region of interest in the centre of a droplet, bleaching with the 488 nm laser line at 4% for 20 s. The photobleached droplet was imaged every 5 s in a format of 512 × 512 pixels at a 12-bit depth (four repetitions). The FRAP analysis was carried out using Fiji software. Before photobleaching, the fluorescence intensity from assigned bleached area and unbleached area was denoted as $I_b^{before}(t)$ and $I_u^{before}(t)$, respectively. $I_b^{after}(t)$ and $I_u^{after}(t)$ represented

the fluorescence intensity from bleach and unbleached area at time t after photobleaching. All fluorescence intensities were normalized by the area. The centre-to-ratio was denoted as $r^{\text{before}} = I_{\text{b}}^{\text{before}}(t)/I_{\text{u}}^{\text{before}}(t)$, $r^{\text{after}} = I_{\text{b}}^{\text{after}}(t)/I_{\text{u}}^{\text{after}}(t)$. Thus, the recovery rate (R) was calculated by

$$R = (r^{\text{after}}(t) - r^{\text{after}}(0))/(r^{\text{before}} - r^{\text{after}}(0)).$$

The recovery rate curve was fitted by an exponential function:

$$R = A(1 - \text{e}^{t/\tau})$$

where A and τ represent the maximal recovery rate and recovery half time, respectively.

Data availability

All data needed to evaluate the conclusions in the paper are present in the paper and/or the supplementary materials. For reproduction purposes, the raw data used to generate the figures are available from the Dryad Digital Repository (DOI:10.6078/D1KH8R).

Code availability

Source code and input scripts supporting this work are available at https://github.com/Shunili/AE-RHP.

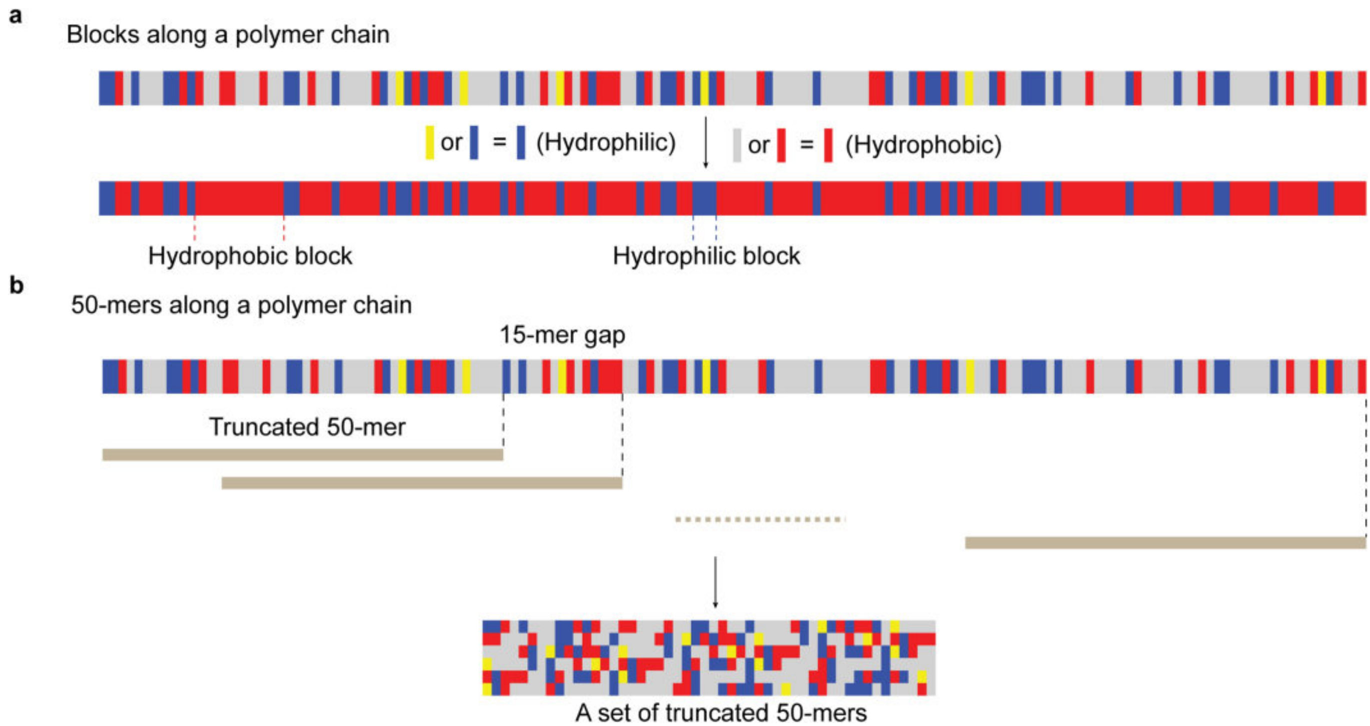
43. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph Model* **14**, 33–38 (1996).
44. AMBER 2019 (Univ. California, San Francisco, 2019).
45. Xiao, L., Zhou, Z. J., Feng, M. L., Tong, A. J. & Xiang, Y. Cationic peptide conjugation enhances the activity of peroxidase-mimicking DNAzymes. *Bioconjugate Chem.* **27**, 621–627 (2016).
46. Comstock, M. J., Ha, T. & Chemla, Y. R. Ultrahigh-resolution optical trap with single-fluorophore sensitivity. *Nat. Methods* **8**, 335–382 (2011).
47. Biewald, L. Machine learning experiment tracking. *Weights & Biases* https://www.wandb.com (2020).

Acknowledgements The work was supported by the US Department of Defense (DOD), Army Research Office, under contract no. W911NF-13-1-0232, Defense Threat Reduction Agency (DTRA) under contract no. HDTRA1-19-1-0011, the National Science Foundation under contract no. DMR-2104443, the US Department of Energy, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division, under contract no. DE-AC02-05-CH11231 (KC3104) and the Alfred P. Sloan Foundation (grant no. G-2021-16757). Z.R. is supported by the Kavli Energy NanoScience Institute through the Kavli ENSI Philomathia Graduate Student Fellowship Program. Scattering studies were done at Advanced Photon Source and use of the Advanced Photon Source was supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences, under contract no. DE-AC02-06CH1135.

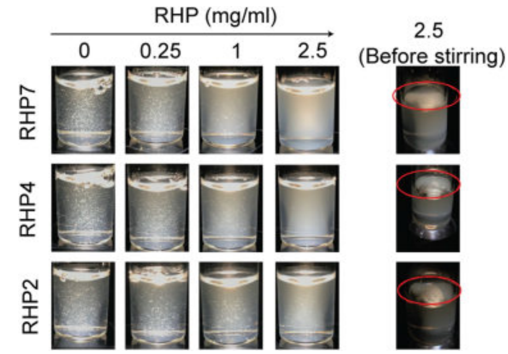
Author contributions T.X. conceived the idea and guided the project. Z.R. and T.J. performed cell-free synthesis of membrane proteins. Z.R. and A.G. performed thermal denaturation of globular enzymes. S.L., I.J., Z.R. and H.H. performed sequence analysis. H.A. and C.B. performed optical tweezers analysis. S.H. and A.A.-K. performed all-atom simulation studies. Z.R. and Z.G. synthesized and characterized the RHPs and DHPs. H.C. performed the cell study. A.G. and H.C. performed the confocal study. All authors participated in writing the manuscript.

Competing interests T.X., H.H., Z.R. and S.L. have a pending PCT patent application. The rest of the authors declare no competing interests.

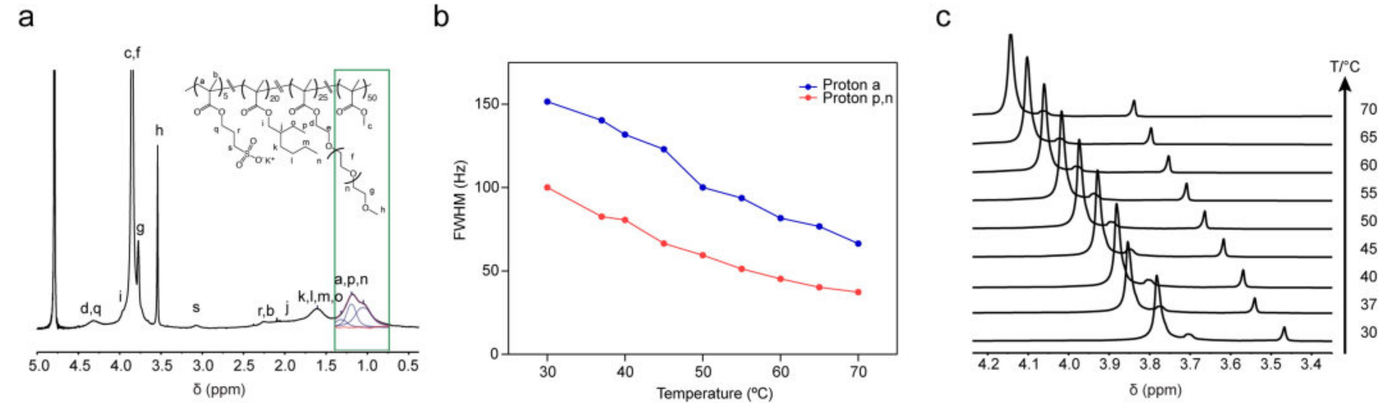
Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-05675-0.
Correspondence and requests for materials should be addressed to Ting Xu.
Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.
Reprints and permissions information is available at http://www.nature.com/reprints.



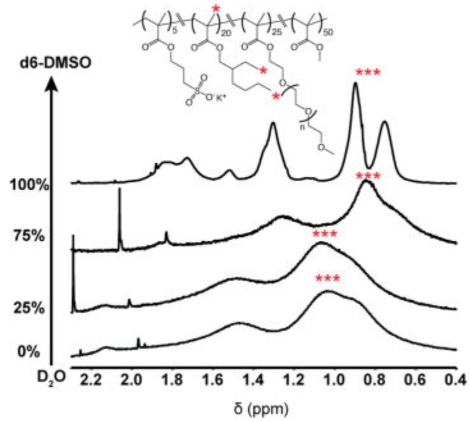
Extended Data Fig. 1 | Blocks and 50-mers along a polymer chain. (a) Each monomer was reassigned to one of two pseudo-monomers (hydrophobic vs. hydrophilic) based on monomer's hydrophobicity. A block comprises consecutive monomers of a single type. (b) A polymer chain was truncated into a set of 50-mers.



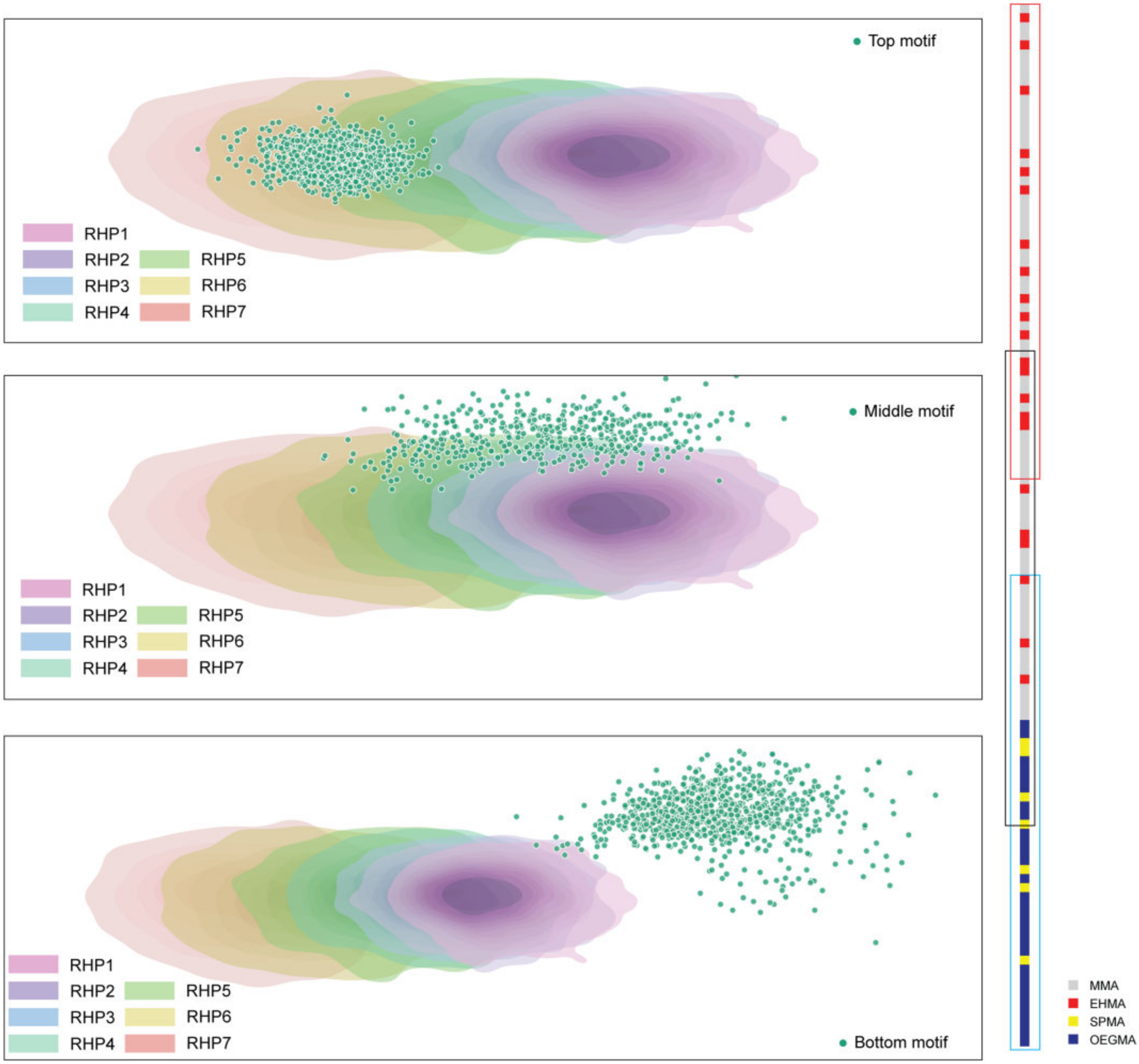
Extended Data Fig. 2 | The FBS solution with different RHP ensembles after incubating at 52 °C for 2.5 h. The FBS solution forms a thin film (like "milk skin") at the air-water interface after thermal treatment (red circle). RHP4 exhibits the weakest tendency for formation of this thin film among three tested RHP ensembles. The solution is stirred to tear the thin film into suspended flakes for visualization.



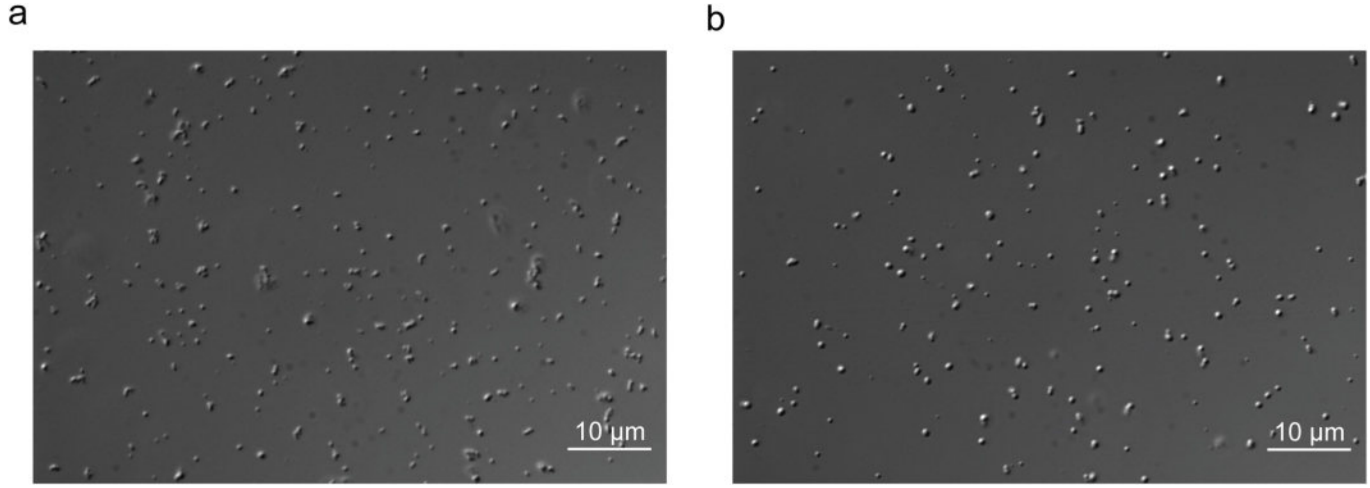
Extended Data Fig. 3 | Temperature-dependent ^{1}H-NMR spectrum of RHP4 in D_2O. (a) ^{1}H-NMR spectrum of RHP4 in D_2O at 37 °C and its chemical structure. (b) The FWHM of proton peaks (a, p, and n) as a function of temperature (30–70 °C). (c) Part of ^{1}H-NMR spectra of RHP4 as a function of temperature (30–70 °C) including proton peaks of the OEGMA side chain.



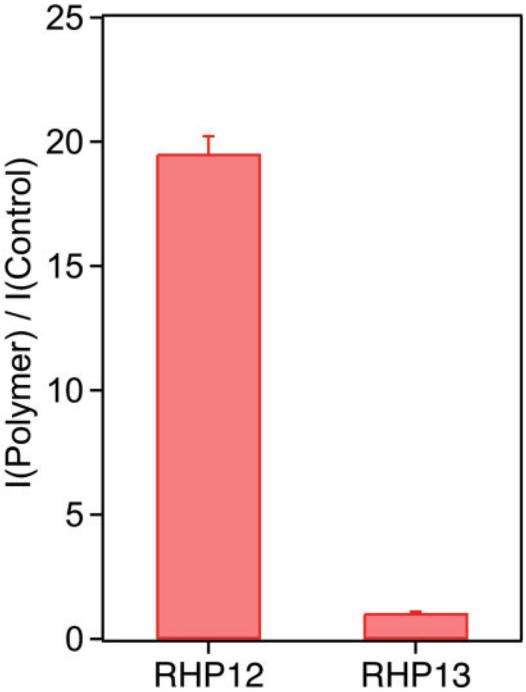
Extended Data Fig. 4 | The effect of solvent polarity on ^{1}H-NMR spectrum of RHP4 in d^{6}-DMSO/D_2O cosolvent.



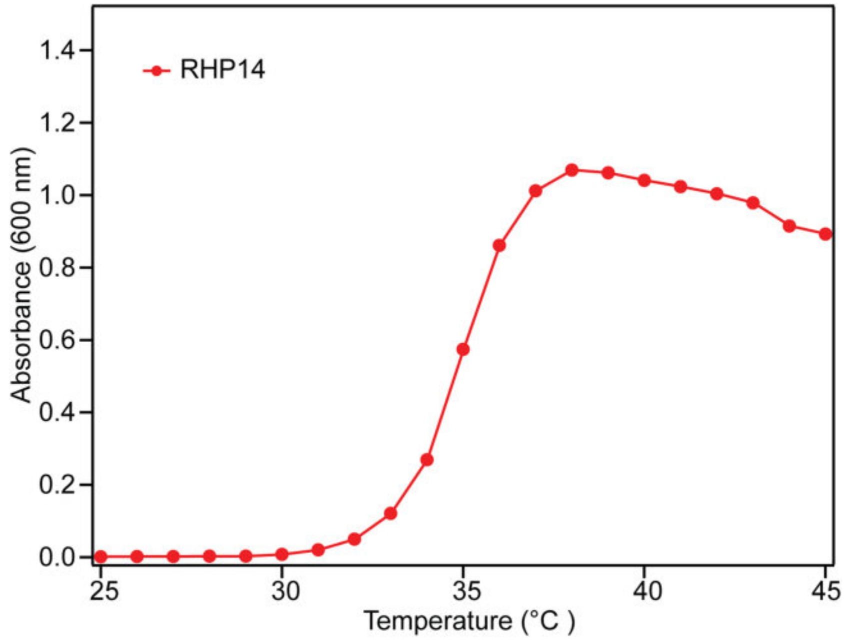
Extended Data Fig. 5 | The distribution of DHP2 segments in PCA space. From top to bottom, the distribution of segments from hydrophobic region, amphiphilic region, and hydrophilic region along DHP2 chains were shown.



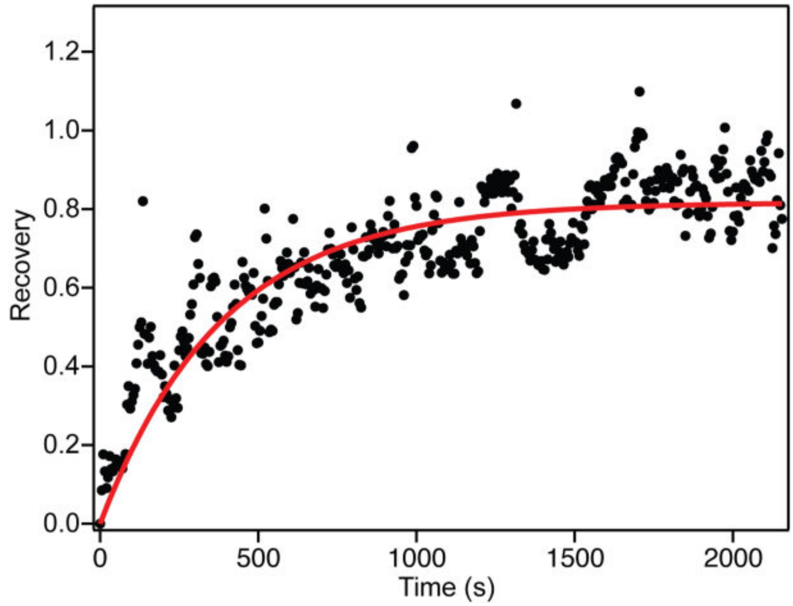
Extended Data Fig. 6 | Differential interference contrast (DIC) images of (a) RHP8 (b) RHP9 phase-separated droplets. Each sample is 1 mg/ml in sodium phosphate buffer (50 mM, pH 7.0).



Extended Data Fig. 7 | Folding status of AqpZ-eGFP in the presence of 0.2 wt% RHPs based on the eGFP fluorescence. Error bar is 1 s.d and n = 3.



Extended Data Fig. 8 | Temperature-dependent turbidimetry for RHP14 ensemble. The polymer solution is 1mg/ml in sodium phosphate buffer (50 mM, pH 7.0).



Extended Data Fig. 9 | FRAP analysis of liquid-like coacervates made from RHP10 ensemble. The recovery trace shows the normalized recovery of a bleached region. Solid red curve fits to an exponential function (see FRAP method section).

Extended Data Table 1 | The conversion from amino acids to RHP monomers (n=4)

ID	Hydrophobicity	Amino acid	Monomer equivalent
1	Hydrophobic	Cys, Tyr, Ala, Thr, and Gly	MMA
2	Hydrophilic	Ser, Gln, His, Asn, and Pro	OEGMA
3	Very hydrophobic	Leu, Ile, Phe, Trp, Val, and Met	EHMA
4	Charged	Glu, Asp, Arg, and Lys	SPMA or DMAEMA

Chlorine activation and enhanced ozone depletion induced by wildfire aerosol

https://doi.org/10.1038/s41586-022-05683-0

Received: 9 August 2022

Accepted: 22 December 2022

Published online: 8 March 2023

Check for updates

Susan Solomon[1,7]✉, Kane Stone[1,7]✉, Pengfei Yu[2], D. M. Murphy[3], Doug Kinnison[4], A. R. Ravishankara[5,6] & Peidong Wang[1]

Remarkable perturbations in the stratospheric abundances of chlorine species and ozone were observed over Southern Hemisphere mid-latitudes following the 2020 Australian wildfires[1,2]. These changes in atmospheric chemical composition suggest that wildfire aerosols affect stratospheric chlorine and ozone depletion chemistry. Here we propose that wildfire aerosol containing a mixture of oxidized organics and sulfate[3–7] increases hydrochloric acid solubility[8–11] and associated heterogeneous reaction rates, activating reactive chlorine species and enhancing ozone loss rates at relatively warm stratospheric temperatures. We test our hypothesis by comparing atmospheric observations to model simulations that include the proposed mechanism. Modelled changes in 2020 hydrochloric acid, chlorine nitrate and hypochlorous acid abundances are in good agreement with observations[1,2]. Our results indicate that wildfire aerosol chemistry, although not accounting for the record duration of the 2020 Antarctic ozone hole, does yield an increase in its area and a 3–5% depletion of southern mid-latitude total column ozone. These findings increase concern[2,12,13] that more frequent and intense wildfires could delay ozone recovery in a warming world.

Massive wildfires in Australia during the austral summer of 2019–2020 (December–January) produced pyrocumulonimbus towers that released about 0.9 Tg of wildfire smoke into the stratosphere[12,14]. Wildfires are also sometimes denoted as bushfires, wildland fires and forest fires; here we refer to them as wildfires. Wildfire smoke can be expected to be primarily composed of organic material, but its stratospheric chemistry is virtually unknown.

Wildfire aerosol composition and ageing

Airborne mass spectrometry data[15–17] showed that carbonaceous compounds were frequently present in about 30–40% of individual particles in the background Northern Hemisphere tropopause and lowermost stratosphere region. The carbonaceous fraction was largely internally mixed with sulfate (that is, both sulfate and carbon were contained within single particles, although whether the latter took the form of a coating was not determined). Although the airborne instrument was not flown through the Australian event, wildfire smoke of the Pacific Northwest Event in 2017 showed about a doubling of the carbonaceous/sulfate aerosol population[17].

Aged stratospheric smoke particles can be characterized by thick coatings[18]. Infrared satellite spectra of the stratospheric wildfire smoke particles after the 2020 event showed signatures of oxidized organic matter, in particular the OH and C=O stretch features[2,19]. The smoke aerosols are referred to hereafter as organic.

Although the detailed composition of stratospheric smoke aerosols has not been determined, tropospheric studies provide insights at lower altitudes. Fresh tropospheric wildfire plumes contain a wide variety of complex organic compounds, including, for example, furans and phenolic compounds, and large-molecular-weight species and humic-like substances can also be present[20]. Levoglucosan is a marker for biomass-burning aerosols derived from the burning of cellulose and hemicellulose and is found in high concentrations in fresh tropospheric plumes[21]. But levoglucosan is oxidized in particles by, for example, reaction with OH, with a lifetime of about 1 day (ref. 22), and secondary aerosol formation by other species is probably also important[23,24]. Here fresh smoke need not be explicitly considered because we are interested in effects over timescales of weeks to months after the fires. The stratosphere is a highly oxidizing environment owing to high concentrations of ozone and free radicals.

There is strong observational evidence for oxidation of many of the complex organic species seen in fresh plumes in the troposphere. Tropospheric observations show a variety of alcohols and acids in aged smoke particles. A wide range of organic acids were identified in tropospheric smoke from Portuguese wildfires, including oleic acid (*cis*-9-octadecenoic acid), succinic acid, heptanedioic acid, malic acid and oxo-acids, as well as several methoxyphenols and alcohols; *n*-alkanols from C_{10} to C_{30} and *n*-alkanoic acids from C_6 to C_{30} were also reported[6,24]. Oxalic acid is frequently among the most abundant single species found in aged tropospheric wildfire smoke aerosols[3–6,25,26]. It has been shown that organic acid content increased with smoke particle ageing and that carbohydrates such as levoglucosan are converted to organic acids during upward transport[3]. Aged aerosols from burning Australian vegetation have also been found to be hygroscopic and contain oxidized material[7], although whether eucalyptus burning might produce different specific organics than other vegetation types merits further study.

[1]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Institute for Environmental and Climate Research, Jinan University, Guangzhou, China. [3]NOAA Chemical Sciences Laboratory, Boulder, CO, USA. [4]Atmospheric Chemistry Observations and Modeling Laboratory, National Center for Atmospheric Research, Boulder, CO, USA. [5]Department of Chemistry, Colorado State University, Fort Collins, CO, USA. [6]Department of Atmospheric Science, Colorado State University, Fort Collins, CO, USA. [7]These authors contributed equally: Susan Solomon, Kane Stone. ✉e-mail: solos@mit.edu; stonek@mit.edu

Whether stratospheric smoke is liquid, glassy or solid must also be considered. Although the Australian plume had a depolarized lidar signature in the stratosphere that generally indicates solid particles, it was partly owing to an unusual size distribution as well as soot in the plume, and can still be consistent with a population of liquid-like particles[27–29]. Several recent papers suggest that organic aerosols may form glasses[30–33]. However, complex mixtures of organics often take on liquid-like properties despite freezing or efflorescence behaviour in the individual components[34,35]. Formation and maintenance of glasses can also be hindered by extra components (such as sulfate and nitrate) and for hygroscopic particles[36]. Following the formalism discussed in Methods, we estimate a diffuso-reactive length scale in organic liquid surfaces on the order of 0.01 microns at mid-latitudes and in polar autumn, driven in large part by the high liquid solubility (hence high particulate concentration) of HCl discussed below. The high HCl abundance in the particle implies that reaction probably occurs very near the surface in particles on the order of a micron in size and reduces the influence of diffusion on the reaction rate. Note that, for stratospheric particles, continuing rapid fluxes of water from/to the surface and the atmosphere as well as the presence of HCl often lead to a quasi-liquid layer, even for solid ice particles (for example, ref. [37]). We suggest that the stratospheric smoke particles probably started out deliquesced as they ascended in the humid environment of dense pyrocumulonimbus clouds, then can be modelled as liquid-like at least for temperatures above about 195 K; we further assume that stratospheric wildfire particles contain large amounts of oxidized organics, particularly organic acids, along with sulfate.

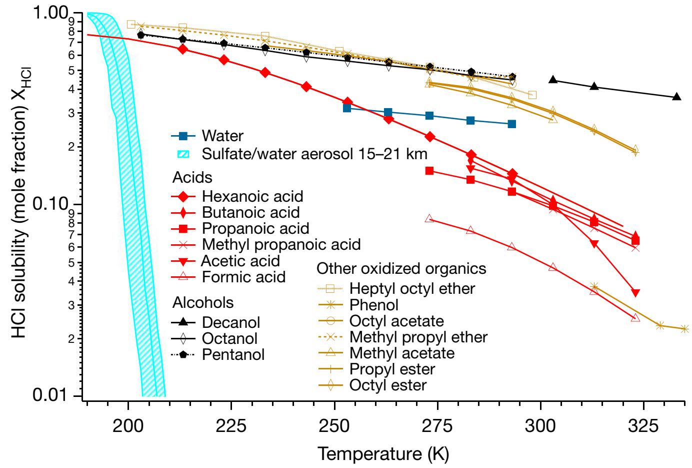


Fig. 1 | HCl solubility in different liquids. Solubility (in mole fraction) of HCl in various organics, pure water and sulfate/water mixtures is shown as a function of temperature based on available laboratory data. The values shown for sulfate/water mixtures are for typical stratospheric conditions over the height range from about 15 to 21 km and are taken from the model used here.

Modelling and testing wildfire chemistry

A sectional aerosol microphysics model (Community Aerosol and Radiation Model for Atmospheres (CARMA)) coupled with the Community Earth System Model (CESM) has been shown to simulate both the observed sulfate and organic/sulfate populations of background lower stratospheric particles[38,39]. The application of this model to the 2020 wildfires reproduced both the observed heating of the stratosphere[12] and effects on NO_x chemistry by means of the $N_2O_5 + H_2O$ heterogeneous reaction[13]. Extended Data Fig. 1 shows that the calculated and observed latitudinal and temporal spread of aerosol extinction during and after the 2020 wildfires are in good general agreement at, for example, 18.5 km; see also ref. [12]. Here we use a specified dynamics version of this state-of-the-art microphysics/atmospheric chemistry model (see Methods). The use of specified dynamics based on observations precludes study of dynamical or radiative feedbacks but allows clear identification of the chemical effects of different chemical processes.

Most of the chlorine released from chlorofluorocarbons resides in HCl and $ClONO_2$ in the lower stratosphere. The influence of organic/sulfate particles on stratospheric chlorine chemistry (such as that of background and volcanic aerosol composed of sulfate/water particles, as well as polar stratospheric clouds (PSCs)) can be expected to be linked to the dissolution and heterogeneous reactions of these species[40]. The most important chlorine processing reactions in/on stratospheric sulfate are

$$ClONO_2 + HCl \rightarrow Cl_2 + HNO_3 \quad (1)$$

$$ClONO_2 + H_2O \rightarrow HOCl + HNO_3 \quad (2)$$

$$HOCl + HCl \rightarrow Cl_2 + H_2O \quad (3)$$

$$HOBr + HCl \rightarrow BrCl + H_2O \quad (4)$$

The rates of these reactions in/on liquid sulfate particles are heavily dependent on temperature and water vapour partial pressure, which control the acidity and hence the solubility of the reactants, particularly HCl (ref. [41]). They are most effective on PSCs in the cold polar region at temperatures below about 195 K (see Fig. 1), at which they take on water and HCl solubility increases. The above heterogeneous reactions can rapidly deplete HCl (and $ClONO_2$), generating ClO and Cl_2O_2 and leading to the Antarctic ozone hole. The observed large 2020 mid-latitude decrease in HCl at much warmer mid-latitude temperatures suggests that HCl dissolution in mixed organic/sulfate smoke should be considered.

There is an extensive literature on HCl solubility in non-aqueous solvents dating to at least the 1930s[42], owing to interest in basic physical chemistry questions such as the role of the dipole moment in the solubility process. Many more studies were published after the 1940s, in part linked to growth in industrial uses of HCl, for example, in syntheses of various plastics and synthetic rubber. Illustrative examples of measured HCl solubility (mole fraction units)[8–11] versus temperature are presented in Fig. 1 for various types of oxidized organic species that may be present in wildfire smoke; pure liquid water and liquid sulfate/water solutions are also shown for reference. The figure demonstrates that HCl is as (or more) soluble in many types of oxidized organic (ranging across alcohols, ethers, esters and acids) than in water at temperatures below about 260 K; it is far more soluble in oxidized organics than in the background pure sulfate/water or PSC particles of the stratosphere unless temperatures fall below about 200 K (for example, in the Antarctic in winter and spring).

The reported solubility of HCl in formic acid (HCOOH) is lower than in the larger organic acids, but the higher-molecular-weight acids (C_3 and greater) show very similar values to one another, with increases at colder temperatures that approach those in the alcohols. The only organic acid for which solubility data below 220 K are available in the published literature is hexanoic acid. Because C_3 and larger oxidized organic acids are probably the main components of aged stratospheric wildfire particles (as they are in the troposphere), here we adopt the solubility of HCl and its temperature dependence in hexanoic acid to approximately represent particulate oxidized organic matter in our numerical model calculations (see Methods). If, for example, more oxidized alcohols or ethers were present in the particles than acids, this would probably increase the HCl solubility slightly.

Because HCl is substantially more soluble in a wide variety of organics at temperatures above about 200 K than it is in background sulfate/water particles based on available laboratory studies (Fig. 1), these data

point towards a transformative role in mid-latitude stratospheric chemistry when oxidized organic aerosols are present. Although alternative elementary reactions are possible, an acid-catalysed mechanism has been proposed[41,43] for reactions (1) and (2):

$$HClONO_2^+(l) + HCl(l) \rightarrow HNO_3(l) + Cl_2(l, g) + H^+ \quad (5)$$

The high solubility of HCl in oxidized organics (Fig. 1) indicates that reaction (5) should occur readily at stratospheric temperatures. This process competes with the first step of reaction (2):

$$HClONO_2^+(l) + H_2O(l) \rightarrow HNO_3(l) + H_2OCl^+(l) \quad (6)$$

The high solubility of HCl in organic acids suggests that the available $HClONO_2^+(l)$ should be more likely to react with HCl(l) than with $H_2O(l)$ in mixed organic/sulfate particles following this mechanism. This is consistent with the known reactivities of reactions (1) and (2) in liquid sulfate aerosols as HCl solubility and hence HCl(l) concentrations increase at cold temperatures below about 195 K, enhancing reaction (1) while suppressing reaction (2); by contrast, reaction (2) is faster than reaction (1) in sulfate particles at warmer temperatures[41]. Reactions (3) and (4) would also be enhanced by high concentrations of HCl(l) and are included here.

To explore the influence of HCl solubility in wildfire particles, we carried out several tests. In one test, we use the HCl solubility in hexanoic acid data (hereafter referred to as solubility case) of Fig. 1 as approximately representative for the oxidized organic/sulfate particles (see Methods), which greatly increases the rates of reactions (1), (3) and (4) at warmer temperatures >200 K. In a second test, we treat the liquid organic portion of the particles like water at all temperatures, that is, we simply impose a dilution factor (dilution case) proportional to the organic content of the organic/sulfate/water particles. This alters the computed H_2SO_4 mole fraction and affects not only reactions (1), (3) and (4) but also reaction (2). This could occur, for example, if the mechanism delineated in reactions (5) and (6) is not valid, enhancing the potential rate of the $ClONO_2 + H_2O$ reaction. In all wildfire model tests, smoke input changes the amount of total aerosol and its associated surface area, enhancing the $N_2O_5 + H_2O$ reaction rate; in a third test, only that process is considered on the wildfire aerosols[13] to compare ozone depletion contributions from different processes (N_2O_5 only case). We compare these three tests to a control run that includes no organic particle chemistry (no organics case).

We find that the model-calculated anomalies in chlorine species for the oxidized organics solubility case are in remarkable agreement overall with the unprecedented extreme changes seen in the 2020 southern mid-latitude observations (Fig. 2). Observed 2020 changes in ozone, HCl, $ClONO_2$ and ClO at 68 hectopascals (hPa) relative to the climatological averages and ranges of available observations for previous years from the Microwave Limb Sounder (MLS) and the Atmospheric Chemistry Experiment (ACE) satellite instruments are presented. Model results are 24-h daily averages, whereas observations shown are averages of available data (sunrise and sunset monthly averages for ACE; averages of daily day and nightside orbits for MLS). Corresponding absolute values for 2020 and the pre-2020 climatologies are shown in Extended Data Fig. 2. HOCl is also reported by ACE but is subject to greater uncertainty, and its anomalies and absolute concentrations are shown in Extended Data Fig. 3. The consistency between the model and measurements obtained across the various species, and over time during the year, strongly supports the proposed mechanism. Differences in absolute abundances probably largely reflect shortcomings in the transport processes of the model.

By contrast, the dilution case greatly overestimates the initial ClO (and HOCl) increases and, accordingly, results in too much ozone loss (Fig. 2). In this case, reaction (2) is enhanced and then declines following

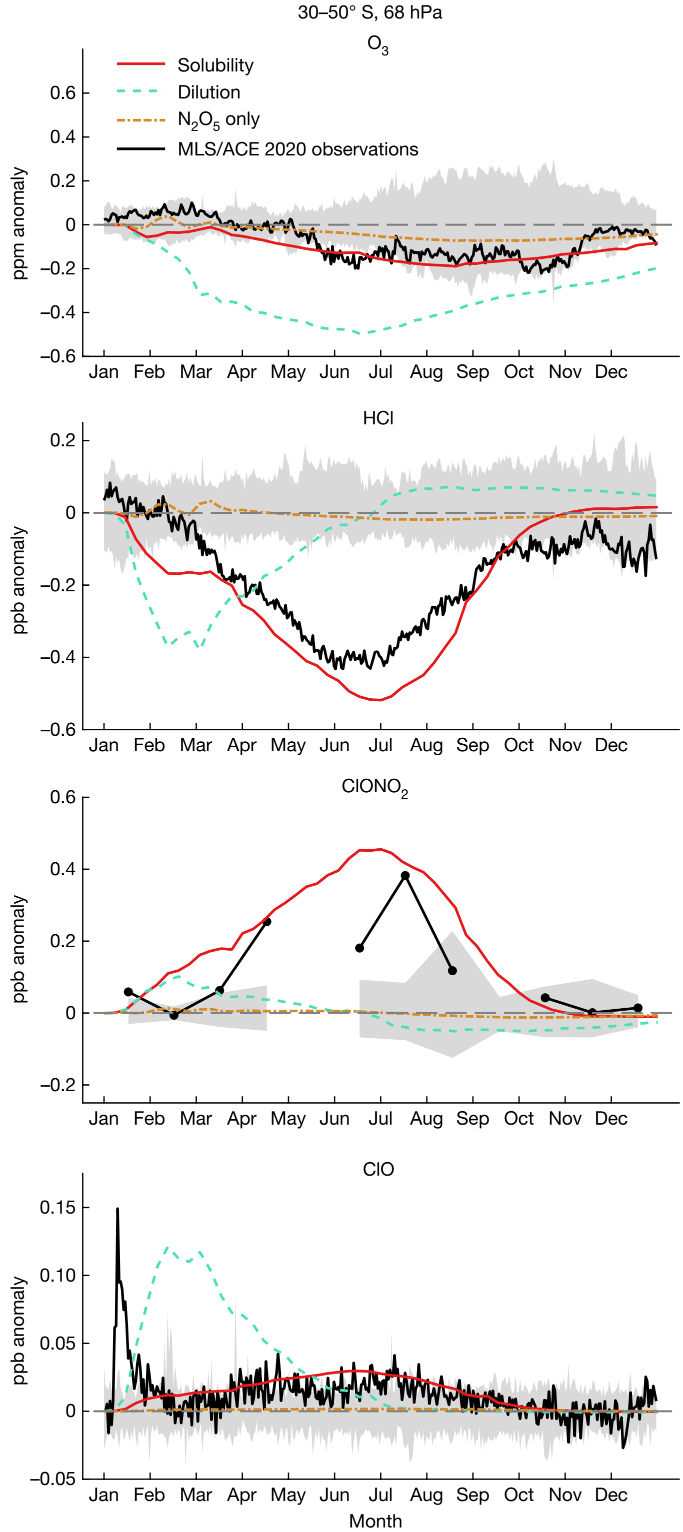


Fig. 2 | Observed and modelled 2020 anomalies in chemical species from 30–50° S at 68 hPa. Grey shaded regions show the ranges of 24-h averaged satellite data anomalies relative to the climatologies of satellite observations before 2020 (daily O_3, HCl and ClO from MLS and monthly $ClONO_2$ from ACE), whereas black lines show the observed anomalies for 2020. Other coloured lines denote calculated anomalies relative to the modelled no organics control for three model test cases: including only N_2O_5 hydrolysis on the aerosols (dashed brown line), considering the added organic material as a dilution factor (dashed green line) and considering the adopted solubility of HCl in organic particles (red line). ppb, parts per billion (by volume throughout this paper); ppm, parts per million (by volume throughout this paper).

the wildfire particle abundances. The fast rate of reaction (2) leads to reduced rather than enhanced autumn $ClONO_2$, as observed.

Taken together, these results indicate that atmospheric HCl indeed dissolves readily in smoke particles under stratospheric conditions, in a manner that is well captured by available solubility data in oxidized organics. This yields subsequent rapid reactions of HCl with $ClONO_2$, HOCl, and HOBr in the particles at relatively warm mid-latitude temperatures, with reaction (1) being the dominant process in the present model. On the basis of the dilution case, it seems that the $ClONO_2 + H_2O$ reaction is suppressed relative to the $ClONO_2$ + HCl reaction in these particles.

In contrast to HCl, satellite measurements indicate that stratospheric nitric acid was not markedly perturbed by the Australian smoke[44] and the model is consistent with those data (Extended Data Fig. 4). High values of $ClONO_2$ are maintained in the sunlit atmosphere in the solubility case, limiting the potential for ozone loss despite a rapid heterogeneous rate for reaction (1). This is owing to continuing HNO_3 photolysis and reaction with OH, leading to reduced but non-zero values of observed NO_2 (ref. [40]) and hence reformation of $ClONO_2$. Thus, a different balance is obtained in mid-latitudes compared with the cold and dark polar regions, in which essentially no NO_2 is available to reform $ClONO_2$ after heterogeneous loss, allowing far greater ClO anomalies to build up and form the ozone hole[40]. Nonetheless, $ClONO_2$ itself photolyses rapidly and is approximately in photochemical steady state with ClO under sunlit mid-latitude conditions, so increased $ClONO_2$ as seen in Fig. 2 implies that some increases in ClO and ozone loss must occur. The shifts in chlorine chemistry result in a threefold calculated increase in ClO, consistent with observations, and this leads to most of the peak local ozone depletion of 10–20% during May to December as shown, which is in accord with the record-low local ozone in this region in June according to the data (Fig. 2). Extended Data Fig. 5 shows that the observed ozone loss profiles on coincident days of observation are similar between ACE and MLS for June–July as an example comparison.

Vertical anomaly profiles of HCl, ClO and $ClONO_2$ averaged over all available data and all model days for June–June 2020 at southern mid-latitudes are also fairly well captured by the solubility case, albeit with some overestimates at higher pressures (lower altitudes), further supporting the proposed mechanism through its links to the profile of wildfire aerosols (Fig. 3). As in Fig. 2, the data do not agree with the model dilution case. Corresponding absolute values are shown in Extended Data Fig. 6.

The model suggests that the wildfire aerosols chemically deplete the mid-latitude total ozone column from 30–50° S by up to 18 Dobson units (DU), roughly triple the amount obtained from N_2O_5 hydrolysis alone, yielding about 3–5% total chemical column loss depending on the month of 2020. Such changes are relatively small compared with the effects of interannual dynamical variability at mid-latitudes and there is evidence for some dynamically driven decreases in ozone in 2020 (refs. [1,45]). Recent studies suggest that the radiative perturbations linked to wildfire smoke affected 2020 Southern Hemisphere dynamical conditions and possibly ozone, which could represent feedbacks in addition to the direct chemical effects evaluated here[46,47]. Nevertheless, the chemical changes are substantial in comparison with the 1% per decade increase expected owing to long-term decreases in halocarbons that have occurred under the international Montreal Protocol and indicate that delays in ozone recovery could occur if wildfires become more frequent or intense in the future.

Some studies have speculated that wildfires may deplete polar ozone[48]. The Antarctic ozone hole of 2020 was both large in area and of record duration (lasting through late December)[49]. In sharp contrast with mid-latitudes, we find that the observed polar 2020 abundances of ozone, HCl, $ClONO_2$ and ClO for 68 hPa from 70–80° S show few marked departures from the ranges of previous years (Fig. 4). The small influence of our wildfire chemistry mechanism on ozone losses in Antarctic spring can be expected because the adopted solubility of HCl in organics becomes smaller than that of typical liquid PSCs below about 200 K (Fig. 1), that is, in polar winter and spring when the ozone hole forms. Larger local ozone changes are obtained close to the tropopause (that is, 100–200 hPa; see Extended Data Fig. 7), at which

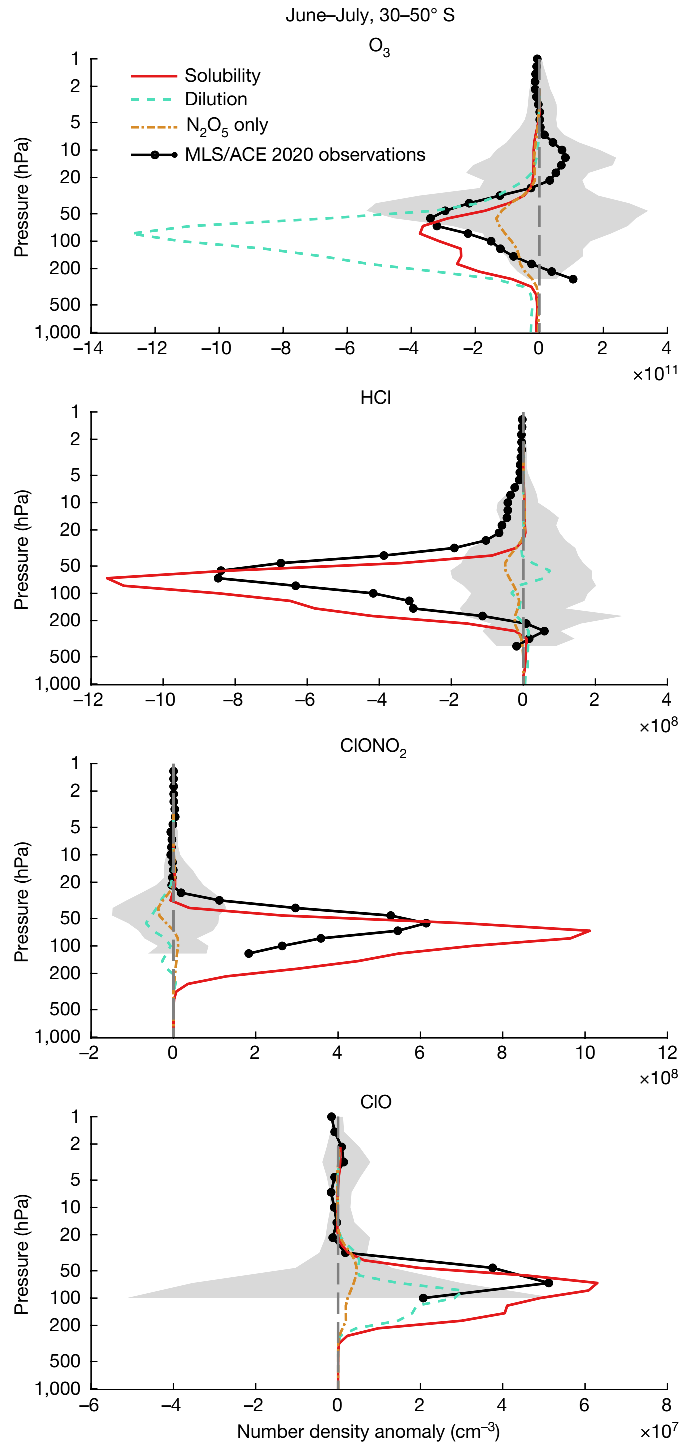


Fig. 3 | Observed and modelled vertical profile anomalies in chemical species from 30–50° S in June–July 2020. Grey shaded regions show the ranges of 24-h averaged satellite data anomalies relative to the climatologies of satellite observations before 2020 (daily O_3 and ClO from MLS and HCl and $ClONO_2$ from ACE, with coverage to low altitudes), whereas black lines show observed anomalies for 2020. Other coloured lines denote calculated anomalies relative to the modelled no organics control for three model test cases: including only N_2O_5 hydrolysis on the aerosols (dashed brown line), considering the added organic material as a dilution factor (dashed green line) and considering the adopted solubility of HCl in organic particles (red line).

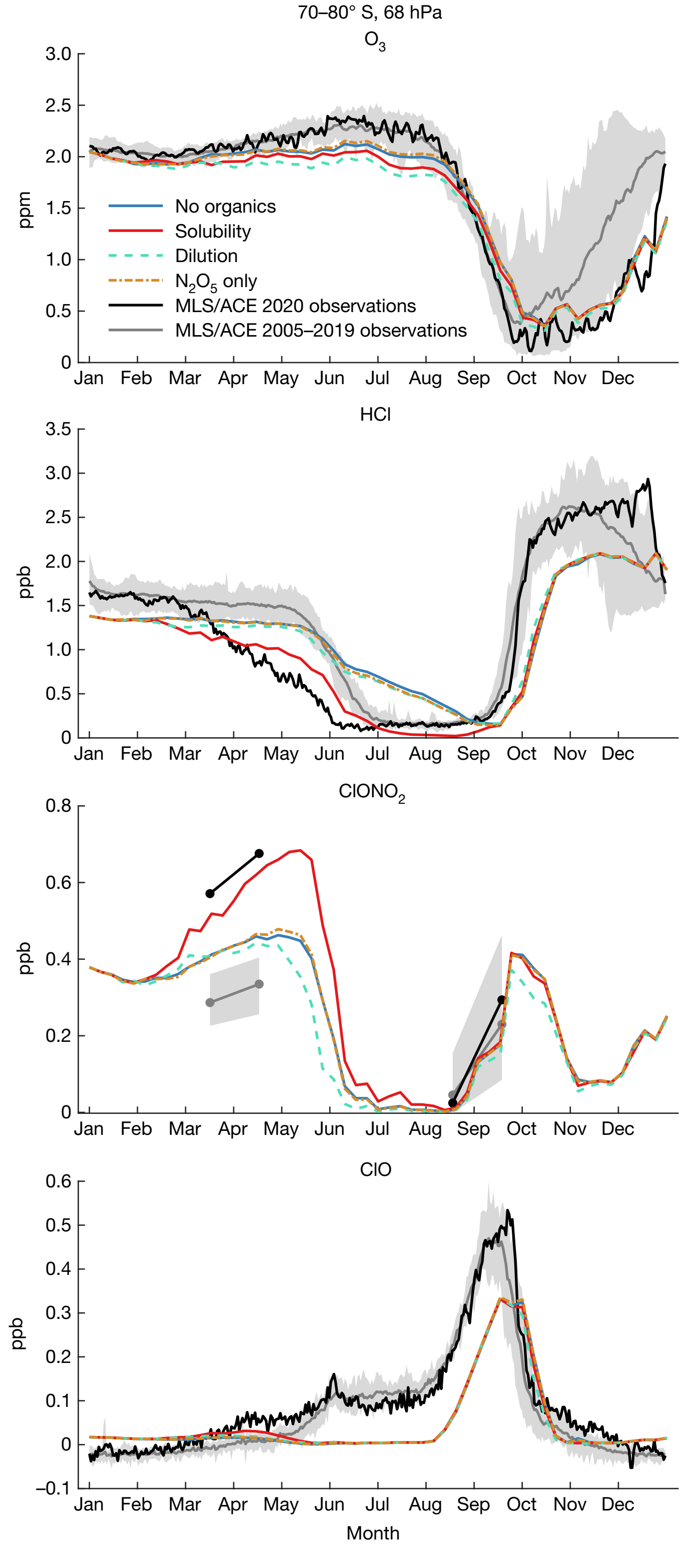


Fig. 4 | Observed and modelled mixing ratios of chemical species at 70–80° S, 68 hPa. Grey shaded regions show the ranges of 24-h averaged satellite data (daily O_3, HCl and ClO from MLS and monthly $ClONO_2$ from ACE) for previous years and the grey line shows their averages, whereas black lines show observations for 2020. Other coloured lines show model-calculated abundances for the no organics control run and for three model test cases: including only N_2O_5 hydrolysis on the aerosols (dashed brown line), considering the added organic material as a dilution factor (dashed green line) and considering the adopted solubility of HCl in organic acid particles (red line).

temperatures are too warm for PSCs. However, the concentration of ozone is relatively small at these altitudes, so the effect on the polar total column depletion is modest (about 5% of the integrated column), much less than that driven by PSCs at higher altitudes.

Notably, observed 2020 polar HCl abundances in the austral autumn season (April–May) decline far earlier (Fig. 4) than in any other year and the oxidized organics solubility case of the model broadly reproduces this unusual behaviour. Temperatures at these latitudes and times are comparable with those in mid-latitudes, so this is to be expected through the added chemistry. The observed HCl inside the polar vortex is also effectively entirely removed by June–July in 2020, considerably lower than the no organics control case or data in other years (Fig. 4 and maps in Extended Data Fig. 8).

A long-standing puzzle in stratospheric chemistry is that the observed timing of early winter HCl decline in the Antarctic and occasionally in the Arctic typically occurs earlier than models predict (see the comprehensive review in ref. [50]), as illustrated in Fig. 4 and Extended Data Fig. 8 for the no organics control case of this model (similar to other models)[50]. Our simulations raise the question of whether early polar winter HCl declines could be linked to background levels of organic particles[15–17] in non-wildfire years.

Although the early winter HCl discrepancy phenomenon is of long-standing chemical interest, our results provide further evidence that it causes limited changes in calculated Antarctic total column ozone depletion in this region[39]. Indeed, the control no organics and organics runs both simulate the 70–80° S 2020 ozone losses and their record long duration well (Fig. 4), implying a dominant role for the observed unusually cold conditions in that year (as imposed on the basis of a reanalysis in these simulations). Inclusion of the wildfire organics HCl solubility does, however, expand the calculated area of the ozone hole (defined as the region in which total column ozone is less than 220 DU) compared with the no organics control by about 2.5 million km^2 in September–October of 2020, contributing to the unusually large ozone hole in that year.

Wildfire aerosol in a warming world

Here we have shown that the effect of wildfire smoke on stratospheric chemistry consists not only of increases in particle surface area (as has been assumed in the published literature)[12,13,45] but, more importantly, in profound impacts on HCl solubility and reactivity. In particular, the 2020 Southern Hemisphere mid-latitude lower stratospheric composition extremes in HCl, $ClONO_2$, ClO and HOCl are remarkably well reproduced by a model that considers the extremely high solubility of HCl in oxidized wildfire smoke organics at warm stratospheric temperatures (as indicated by historical laboratory solubility data) and subsequent reactions. These in turn deplete mid-latitude ozone.

A record-early observed 2020 southern autumn (April–May) disappearance of HCl at 70–80° S is another finding of this paper and is well simulated by this model. Whether the smaller amounts of organic particles present in other years may be responsible for discrepancies between calculated and observed HCl declines[50] in high-latitude autumn and winter documented in the published literature merits future study.

Although the model indicates that the organic particles expanded the size of the September–October 2020 ozone hole by about 2.5 million km^2, it does not explain its record longevity[49], raising the question of the roles of natural variability and possible forcings (for example, relation to the unusual mid-latitude ozone losses as well as the radiative effects of the particles themselves and resulting impacts on temperatures, winds and stratosphere–troposphere coupling[46,47,49]). Further work is needed to examine potential radiative and dynamical feedbacks of the aerosol and ozone changes.

The same chemistry discussed here for the southern mid-latitudes and polar regions can also be expected in Northern Hemisphere wildfire smoke. Further, similar reactions should occur wherever aged organic

aerosols and HCl are found, that is, not only in wildfire smoke but also in biomass burning, pollution and, potentially, aircraft-generated aerosols, both in the lower stratosphere and in the troposphere. Consequences of HCl solubility in oxidized organics for long-term lower stratospheric ozone trends also merit further study, along with potentially altered solubilities of other compounds besides HCl in such particles in tropospheric and stratospheric chemistry.

In closing, limitations of our chemical assumptions should be noted. For example, it is possible that the organic material could freeze, even doing so at warmer temperatures than normal, and thereby prolonging heterogeneous chemistry in polar spring (and possibly autumn). Other reactions beyond those considered may also be important, and new laboratory studies of solubility and reactions of stratospheric species in oxidized organic aerosols and especially wildfire aerosols are desperately needed if, as expected, wildfire frequency and intensity increase in a warming world.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-05683-0.

1. Santee, M. L. et al. Prolonged and pervasive perturbations in the composition of the Southern Hemisphere midlatitude lower stratosphere from the Australian New Year's fires. *Geophys. Res. Lett.* **49**, e2021GL096270 (2022).
2. Bernath, P., Boone, C. & Crouse, J. Wildfire smoke destroys stratospheric ozone. *Science* **375**, 1292–1295 (2022).
3. Gao, S. et al. Water-soluble organic components in aerosols associated with savanna fires in Southern Africa: identification, evolution, and distribution. *J. Geophys. Res.* **108**, 8491 (2003).
4. Kundu, S., Kawamura, K., Andreae, T. W., Hoffer, A. & Andreae, M. O. Molecular distributions of dicarboxylic acids, ketocarboxylic acids and α-dicarbonyls in biomass burning aerosols: implications for photochemical production and degradation in smoke layer. *Atmos. Chem. Phys.* **10**, 2209–2225 (2010).
5. Trebs, I. et al. Real-time measurements of ammonia, acidic trace gases and water-soluble inorganic aerosol species at a rural site in the Amazon Basin. *Atmos. Chem. Phys.* **4**, 967–987 (2004).
6. Vicente, A. et al. Organic speciation of aerosols from wildfires in central Portugal during summer 2009. *Atmos. Environ.* **57**, 186–196 (2012).
7. Mallet, M. D. et al. Composition, size and cloud condensation nuclei activity of biomass burning aerosol from northern Australian savannah fires. *Atmos. Chem. Phys.* **17**, 3605–3617 (2017).
8. Ahmed, W., Gerrard, W. & Malukdar, V. K. Significance of the solubility of hydrogen halides in liquid compounds. *J. Appl. Chem.* **20**, 109–116 (1970).
9. Gerrard, W. & Macklen, E. Solubility of hydrogen halides in organic compounds containing oxygen. I. Solubility of hydrogen chloride in alcohols, carboxylic acids and esters. *J. Appl. Chem.* **6**, 241–244 (1956).
10. Gerrard, W., Mincer, A. M. A. & Wyvill, P. L. Solubility of hydrogen halides in organic compounds containing oxygen. III. Solubility of hydrogen chloride in alcohols and certain esters at low temperatures. *J. Appl. Chem.* **9**, 89–93 (1959).
11. International Union of Pure and Applied Chemistry (IUPAC), Fogg, P. G. T., Gerrard, W. & Clever, H. L. (eds) *Hydrogen Halides in Non-aqueous Solvents (Solubility Data Series)* Vol. 42 (Pergamon, 1990).
12. Yu, P. et al. Persistent stratospheric warming due to 2019–2020 Australian wildfire smoke. *Geophys. Res. Lett.* **48**, e2021GL092609 (2021).
13. Solomon, S. et al. On the stratospheric chemistry of midlatitude wildfire smoke. *Proc. Natl Acad. Sci. USA* **119**, e2117325119 (2022).
14. Peterson, D. A. et al. Australia's Black Summer pyrocumulonimbus super outbreak reveals potential for increasingly extreme stratospheric smoke events. *NPJ Clim. Atmos. Sci.* **4**, 38 (2021).
15. Murphy, D. M., Thomson, D. S. & Mahoney, M. J. In situ measurements of organics, meteoritic material, mercury, and other elements in aerosols at 5 to 19 kilometers. *Science* **282**, 1664–1669 (1998).
16. Murphy, D. M., Cziczo, D. J., Hudson, P. K. & Thomson, D. S. Carbonaceous material in aerosol particles in the lower stratosphere and tropopause region. *J. Geophys. Res.* **112**, D04203 (2007).
17. Murphy, D. M. et al. Radiative and chemical implications of the size and composition of aerosol particles in the existing or modified global stratosphere. *Atmos. Chem. Phys.* **21**, 8915–8932 (2021).
18. Ditas, J., Ma, N., Zhang, Y. & Cheng, Y. Strong impact of wildfires on the abundance and aging of black carbon in the lowermost stratosphere. *Proc. Natl Acad. Sci. USA* **115**, E11595–E11603 (2018).
19. Boone, C. D., Bernath, P. F. & Fromm, M. D. Pyrocumulonimbus stratospheric plume injections measured by the ACE-FTS. *Geophys. Res. Lett.* **47**, e2020GL088442 (2020).
20. Palm, B. B., Peng, Q., Fredrickson, C. D. & Thornton, J. A. Quantification of organic aerosol and brown carbon evolution in fresh wildfire plumes. *Proc. Natl Acad. Sci. USA* **117**, 29469–29477 (2020).
21. Sannigrahi, P., Sullivan, A. P., Weber, R. J. & Ingall, E. D. Characterization of water-soluble organic carbon in urban atmospheric aerosols using solid-state ^{13}C NMR spectroscopy. *Environ. Sci. Technol.* **40**, 666–672 (2006).
22. Garofalo, L. A. et al. Emission and evolution of submicron organic aerosol in smoke from wildfires in the western United States. *ACS Earth Space Chem.* **3**, 1237–1247 (2019).
23. Cappa, C. D. et al. Biomass-burning-derived particles from a wide variety of fuels – part 2: effects of photochemical aging on particle optical and chemical properties. *Atmos. Chem. Phys.* **20**, 8511–8532 (2020).
24. Vicente, A. et al. Emission factors and detailed chemical composition of smoke particles from the 2010 wildfire season. *Atmos. Environ.* **71**, 295–303 (2013).
25. Mochida, M. et al. Spatial distributions of oxygenated organic compounds (dicarboxylic acids, fatty acids, and levoglucosan) in marine aerosols over the western Pacific and off the coast of East Asia: continental outflow of organic aerosols during the ACE-Asia campaign. *J. Geophys. Res. Atmos.* **108**, 8638 (2003).
26. Deshmukh, D. K. et al. High loadings of water-soluble oxalic acid and related compounds in PM2.5 aerosols in eastern central India: influence of biomass burning and photochemical processing. *Aerosol Air Qual. Res.* **19**, 2625–2644 (2019).
27. Haarig, M. et al. Depolarization and lidar ratios at 355, 532, and 1064 nm and microphysical properties of aged tropospheric and stratospheric Canadian wildfire smoke. *Atmos. Chem. Phys.* **18**, 11847–11861 (2018).
28. Ohneiser, K. et al. Smoke of extreme Australian bushfires observed in the stratosphere over Punta Arenas, Chile, in January 2020: optical thickness, lidar ratios, and depolarization ratios at 355 and 532 nm. *Atmos. Chem. Phys.* **20**, 8003–8015 (2020).
29. Ansmann, A. et al. Tropospheric and stratospheric wildfire smoke profiling with lidar: mass, surface area, CCN, and INP retrieval. *Atmos. Chem. Phys.* **21**, 9779–9807 (2021).
30. Zobrist, B., Marcolli, C., Pedenera, D. A. & Koop, T. Do atmospheric aerosols form glasses? *Atmos. Chem. Phys.* **8**, 5221–5244 (2008).
31. Virtanen, A. et al. An amorphous solid state of biogenic secondary organic aerosol particles. *Nature* **467**, 824–827 (2010).
32. Reid, J. P. et al. The viscosity of atmospherically relevant organic particles. *Nat. Commun.* **9**, 956 (2018).
33. Boomian, V. et al. Sunlight can convert atmospheric aerosols into a glassy solid state and modify their environmental impacts. *Proc. Natl Acad. Sci. USA* **119**, e2208121119 (2022).
34. Cappa, C. D., Lovejoy, E. R. & Ravishankara, A. R. Evidence for liquid-like and nonideal behavior of a mixture of organic aerosol components. *Proc. Natl Acad. Sci. USA* **105**, 18687–18691 (2008).
35. Marcolli, C., Luo, B. P. & Peter, T. Mixing of the organic aerosol fractions: liquids as the thermodynamically stable phases. *J. Phys. Chem. A* **108**, 2216–2224 (2004).
36. Koop, T., Bookhold, J., Shirawa, M. & Poeschl, U. Glass transition and phase state of organic compounds: dependency on molecular properties and implications for secondary organic aerosols in the atmosphere. *Phys. Chem. Chem. Phys.* **13**, 19238–19255 (2011).
37. McNeill, V. F., Loerting, T., Geiger, F. M., Trout, B. L. & Molina, M. J. Hydrogen chloride-induced surface disordering on ice. *Proc. Natl Acad. Sci. USA* **103**, 9422–9427 (2006).
38. Yu, P. et al. Evaluations of tropospheric aerosol properties simulated by the community earth system model with a sectional aerosol microphysics scheme. *J. Adv. Model. Earth Syst.* **7**, 865–914 (2015).
39. Yu, P. et al. Radiative forcing from anthropogenic sulfur and organic emissions reaching the stratosphere. *Geophys. Res. Lett.* **43**, 9361–9367 (2016).
40. Solomon, S. Stratospheric ozone depletion: a review of concepts and history. *Rev. Geophys.* **37**, 275–316 (1999).
41. Shi, Q., Jayne, J. T., Kolb, C. E., Worsnop, D. R. & Davidovits, P. Kinetic model for reaction of $ClONO_2$ with H_2O and HCl and HOCl with HCl in sulfuric acid solutions. *J. Geophys. Res.* **106**, 24259–24274 (2001).
42. Bell, R. P., The electrical energy of dipole molecules in solution, and the solubilities of ammonia, hydrogen chloride, and hydrogen sulphide, in various solvents. *J. Chem. Soc.* https://doi.org/10.1039/JR9310001371 (1931).
43. Robinson, G. N., Worsnop, D. R., Jayne, J. T., Kolb, C. E. & Davidovits, P. Heterogeneous uptake of $ClONO_2$ and N_2O_5 by sulfuric acid solutions. *J. Geophys. Res.* **102**, 3583–3601 (1997).
44. Schwartz, M. J. et al. Australian New Year's pyroCb Impact on stratospheric composition. *Geophys. Res. Lett.* **47**, e2020GL090831 (2020).
45. Strahan, S. E. et al. Unexpected repartitioning of stratospheric inorganic chlorine after the 2020 Australian wildfires. *Geophys. Res. Lett.* **49**, e2022GL098290 (2022).
46. Yook, S., Thompson, D. W. J. & Solomon, S. Climate impacts and potential drivers of the unprecedented Antarctic ozone holes of 2020 and 2021. *Geophys. Res. Lett.* **49**, e2022GL098064 (2022).
47. Damany-Pearce, L. et al. Australian wildfires cause the largest stratospheric warming since Pinatubo and extends the lifetime of the Antarctic ozone hole. *Sci. Rep.* **12**, 12665 (2022).
48. Ohneiser, K. et al. The unexpected smoke layer in the High Arctic winter stratosphere during MOSAiC 2019–2020. *Atmos. Chem. Phys.* **21**, 15783–157808 (2021).
49. Klekociuk, A. R. et al. The Antarctic ozone hole during 2020. *J. South. Hemisph. Earth Syst. Sci.* **72**, 19–37 (2022).
50. Grooß, J.-U. et al. On the discrepancy of HCl processing in the core of the wintertime polar vortices. *Atmos. Chem. Phys.* **18**, 8647–8666 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2023

Methods

Model simulations

We use the sectional aerosol model CARMA coupled with the National Science Foundation/Department of Energy CESM (CESM-CARMA)[38,39,51,52]. We spin-up by running the model from a multiyear free-running simulation and continue in free-running mode from 29 December 2019 to 1 March 2020, which allows the initial injection to be self-lofted into the stratosphere, as shown in ref. [12]. We inject 0.9 Tg of smoke (three times that of the Pacific Northwest fires in 2017) during the fire days of 29–31 December 2019 and 4 January 2020 over southeastern Australia (39° S, 150° E) at an altitude of 12 km (ref. [12]). The smoke is made up of 2.5% black carbon, which is a factor in the lofting of the injected smoke; sensitivity simulations with other values in this model[12] show that this percentage gave the best match to satellite observations in this model. After 1 March, simulated winds and temperatures are nudged to the Goddard Earth Observing System version 5 analysis (GEOS-5)[53] to ensure accurate meteorological conditions for 2020 that are insensitive to the upper boundary and optimize comparisons with observations.

The model includes 56 layers from the surface up to 1.8 hPa (about 45 km), with a vertical resolution of about 1 km in the upper troposphere and lower stratosphere. The horizontal resolution of the model is 1.9° latitude × 2.5° longitude.

The model calculates two kinds of organic aerosols: (1) organic material mixed with sulfate, black carbon, sea salt and dust and (2) purely organic smoke particles emitted from the fires. Both sets of particles are computed in 20 size bins whose radii range from 0.05 to 8.7 μm. The model includes both the oxidation of the organic component of the smoke by ozone and condensation of sulfuric acid on the smoke particles[12].

Stratospheric heterogeneous chemistry rates in and on aerosols are calculated using the kinetics parameterization approach generally used in atmospheric chemistry models[41] for the heterogeneous reactions of $ClONO_2$ + HCl, $ClONO_2$ + H_2O and HOCl + HCl. The kinetics approach uses experimental data for HCl, HOCl and $ClONO_2$ liquid diffusion, solubility and reactivity to obtain the heterogeneous reactions for H_2SO_4/H_2O stratospheric aerosol heterogeneous kinetics.

To investigate the effect of smoke on stratospheric chemistry, four model setups are used that adopt different heterogeneous chemistry on stratospheric aerosols. Further detail on the oxidized organic HCl solubility and dilution cases (see main text) is given in the following.

Dilution. This simulation assumes that the organic carbon behaves as an aerosol diluent and combines with the H_2SO_4/H_2O aerosols, which effectively lowers the H_2SO_4 weight percent that is a key parameter in the kinetics parameterization[41]. This is achieved by

$$\mathrm{wt}_{\mathrm{d}} = \frac{1}{\frac{1}{\mathrm{wt}} + \frac{\mathrm{OC}}{\mathrm{SO_4}}}$$

in which wt is the weight percent of H_2SO_4 in the H_2SO_4/H_2O solution, $\mathrm{wt_d}$ is the diluted H_2SO_4 weight percent, OC is the organic carbon mass concentration and SO_4 is the sulfur mass concentration. An estimate of the organic molar mass of 116 (hexanoic acid) is used to obtain the organic-weight-corrected H_2SO_4 mole fraction as follows

$$x_{\mathrm{H_2SO_4}} = \frac{\mathrm{wt_d}}{\mathrm{wt_d} + \frac{\mathrm{wt_{H_2O}} \times 98}{18} + \frac{\mathrm{wt_{OC}} \times 98}{116}}$$

which is then used in the calculation of HCl solubility. Here $\mathrm{wt_{H_2O}}$ and $\mathrm{wt_{OC}}$ are the weight percent of water and organic carbon, respectively, 98 is the molar mass of H_2SO_4, 18 is the molar mass of H_2O and 116 is the molar mass of hexanoic acid.

Solubility. This simulation alters the HCl solubility to account for the effects of smoke organics using laboratory measurements[8–11] (instead of changing the H_2SO_4 weight percent). This is done for all locations and times when the ratio of mass concentration of organic carbon to sulfur is >1; otherwise, the original parameterization[41] is used. Following ref. [11], the HCl mole fraction in hexanoic acid is calculated using the following fit function to the laboratory data points (see red diamonds in Fig. 1, in which the red line through them indicates this fit):

$$x_{\mathrm{HCl}} = \exp\left(28.99 - \frac{3{,}300.46}{T} - 18.14 \times \log\left(\frac{T}{100}\right)\right)$$

in which T is the temperature. As hexanoic acid is a weak acid, we assume that HCl dissociates in hexanoic acid similarly to water. Mole fraction is converted to mole ratio (r_{HCl}) and used to calculate an effective Henry's law coefficient using a dissociation constant of HCl in water of $10^{5.9}$ (following ref. [54])

$$H_{\mathrm{HCl}} = \frac{r_{\mathrm{hcl}} \times \rho_{\mathrm{hex}} \times 10^{5.9}}{116}$$

in which ρ_{hex} is given by[55]:

$$\rho_{\mathrm{hex}} = (-5.01 \times 10^{-7} \times T^2 - 5.23 \times 10^{-4} \times T + 1.12) \times 1{,}000$$

The rates of the heterogeneous reactions are then calculated in the kinetics parameterization[41] using this new value of Henry's law for HCl solubility in hexanoic acid. Because organic acids are weak acids, we assume that the acidity in a mixed organic/sulfuric acid solution would be determined largely by the sulfuric acid content. We do not consider the possible increase in pH owing to interactions of these organics with sulfuric acid and assume such changes to be small.

A caveat of this approach is that, in principle, the effective Henry's law best applies to dilute solutions so that some uncertainty occurs for strong solutions but laboratory data suggest that this is a small effect for our purposes[8].

Observations

We use the MLS Level 2, version 5 PressureZM measurement product for O_3, HCl, HNO_3 and ClO data[56] to analyse mid-latitude (30–50° S) and polar (70–80° S) time series at 68 hPa as well as O_3 and ClO vertical profiles between 30° S and 50° S. We use the Level 2, version 5 PressureGrid measurement product for HCl maps. MLS has the benefit of daily measurements with few data gaps and allows for comprehensive temporal comparison with the model over all months in 2020. However, note that the ClO measurements at lower levels (68–147 hPa) are subject to a known negative bias. At 68 hPa, the bias is relatively small and has been reduced in version 5 compared with the previous version 4.2. It is therefore recommended[56] that day minus night measurements be used to reduce the bias. However, for the purposes of this paper in which comparison with model weekly average output is required and absolute values are less important as opposed to anomalies, we opted to use the averages of MLS day and night data. This also has the extra benefit of not having to account for further biases in the polar region for the MLS day minus night product.

Also, we use the ACE-Fourier transform spectrometer (FTS) version 4.1 data[57] for $ClONO_2$ and HOCl time series at 68 hPa and between 30° S and 50° S for mid-latitudes and 70° S and 80° S for polar conditions. We also use ACE for vertical profiles of $ClONO_2$ and HCl during June and July between 30° S and 50° S. The use of ACE HCl data here maximizes altitude coverage. ACE has sporadic spatial coverage for specific latitude ranges. Therefore, for the time series of $ClONO_2$, monthly averages of the available daily data for each month are constructed. For example, there is no data coverage between 30° S and 50° S in May and September. ACE quality control is performed by removing data points that lie outside three standard deviations of the mean values, as suggested for this data product[57].

ACE and model data are interpolated onto a regular grid that coincides with the MLS pressure grid product to allow for appropriate comparison between datasets.

Data availability

All data used in this study are publicly available. MLS data: https://disc.gsfc.nasa.gov/datasets?page=1&source=Aura%20MLS; ACE-FTS data: http://www.ace.uwaterloo.ca (with registration: https://databace.scisat.ca/l2signup.php); CESM1-CARMA: https://doi.org/10.7910/DVN/GHNJQA.

Code availability

The model used in this study can be accessed at https://www2.cesm.ucar.edu/models/cesm1.2/cesm/doc/usersguide/x290.html. The changes described herein for the kinetics parameterization are available at https://doi.org/10.7910/DVN/GHNJQA.

51. Bardeen, C. G., Toon, O. B., Jensen, E. J., Marsh, D. R. & Harvey, V. L. Numerical simulations of the three-dimensional distribution of meteoric dust in the mesosphere and upper stratosphere. *J. Geophys. Res.* **113**, D17202 (2008).
52. Toon, O. B., Turco, R. P., Westphal, D., Malone, R. & Liu, M. A multidimensional model for aerosols: description of computational analogs. *J. Atmos. Sci.* **45**, 2123–2143 (1988).
53. Rienecker, M. M. et al. *Technical Report Series on Global Modeling and Data Assimilation: The GEOS-5 Data Assimilation System-Documentation of Versions 5.0.1, 5.1.0, and 5.2.0* Vol. 27 (National Aeronautics and Space Administration, Goddard Space Flight Center, 2008).
54. Trummal, A., Lipping, L., Kaljurand, I., Koppel, I. A. & Leito, I. Acidity of strong acids in water and dimethyl sulfoxide. *J. Phys. Chem. A* **120**, 3663–3669 (2016).
55. Ghatee, M. H., Ghanavati, F., Bahrami, M. & Zolghadr, A. R. Molecular dynamics simulation and experimental approach to the temperature dependent surface and bulk properties of hexanoic acid. *Ind. Eng. Chem. Res.* **52**, 3334–3341 (2013).
56. Livesey, N. J. et al. Version 5.0x Level 2 and 3 data quality and description document. *Jet Propulsion Laboratory* http://mls.jpl.nasa.gov (2020).
57. Boone, C. D., Bernath, P. F., Cok, D., Jones, S. C. & Steffen, J. Version 4 retrievals for the atmospheric chemistry experiment Fourier transform spectrometer (ACE-FTS) and imagers. *J. Quant. Spectrosc. Radiat. Transf.* **247**, 106939 (2020).

Acknowledgements S.S. and K.S. are partly supported by NSF 1848863. D.K. was financed in part by NASA grant 80NSSC19K0952. P.Y. is supported by the National Natural Science Foundation of China (42175089, 42121004). D.M.M. is supported by NOAA base and climate funding. The CESM project is supported by the National Science Foundation and the Office of Science (BER) of the U.S. Department of Energy. We gratefully acknowledge high-performance computing support from Cheyenne (https://doi.org/10.5065/D6RX99HX) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation.

Author contributions S.S. and K.S. contributed equally and are co-first authors of this study. S.S., K.S., P.Y. and D.M.M. designed the initial work. K.S. analysed the data and refined the study design and produced the figures. S.S. drafted the initial text. K.S., D.M.M., D.K., A.R.R. and P.W. contributed substantially to the interpretation of findings and to the revisions of the manuscript.

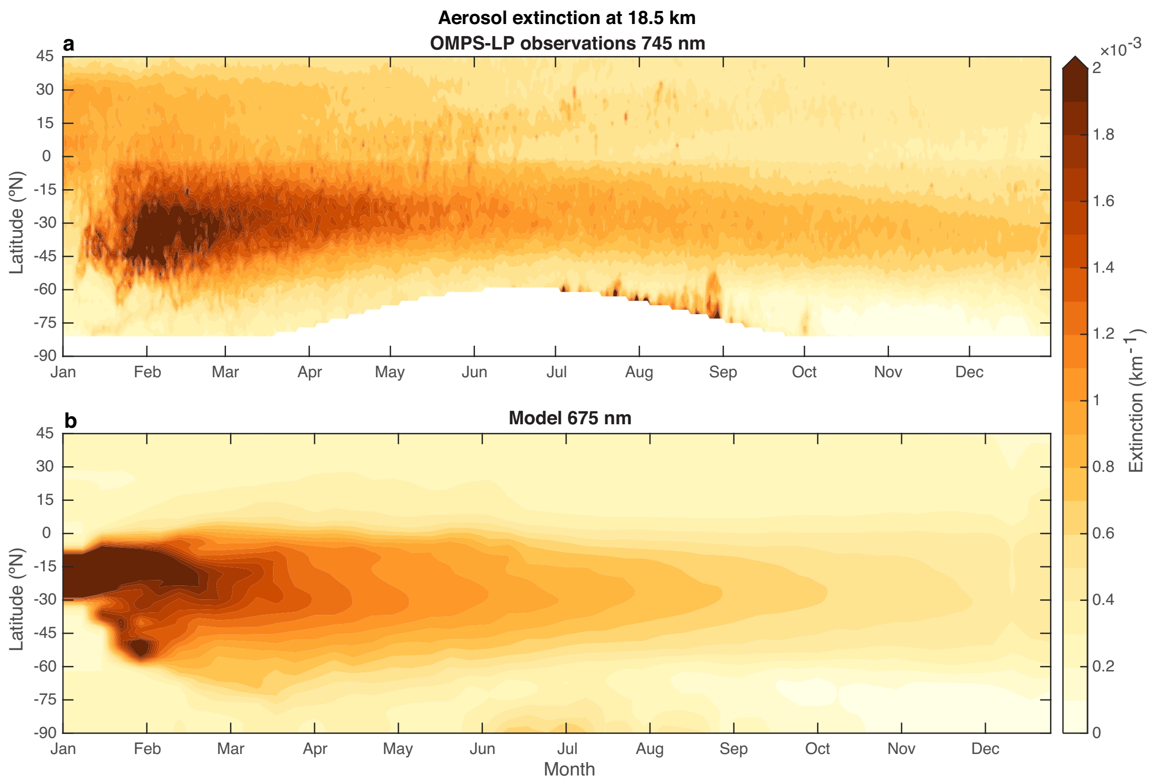
Competing interests The authors declare no competing interests.

Additional information

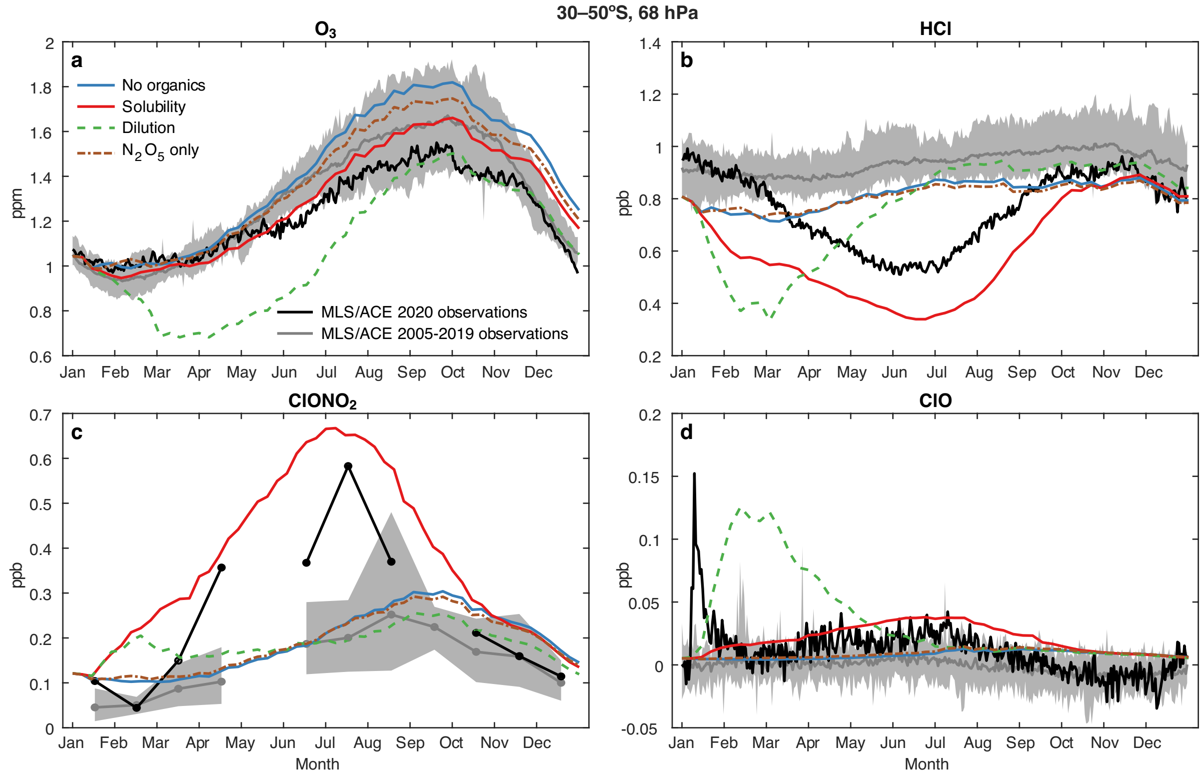
Correspondence and requests for materials should be addressed to Susan Solomon or Kane Stone.

Peer review information *Nature* thanks Clare Paton-Walsh and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

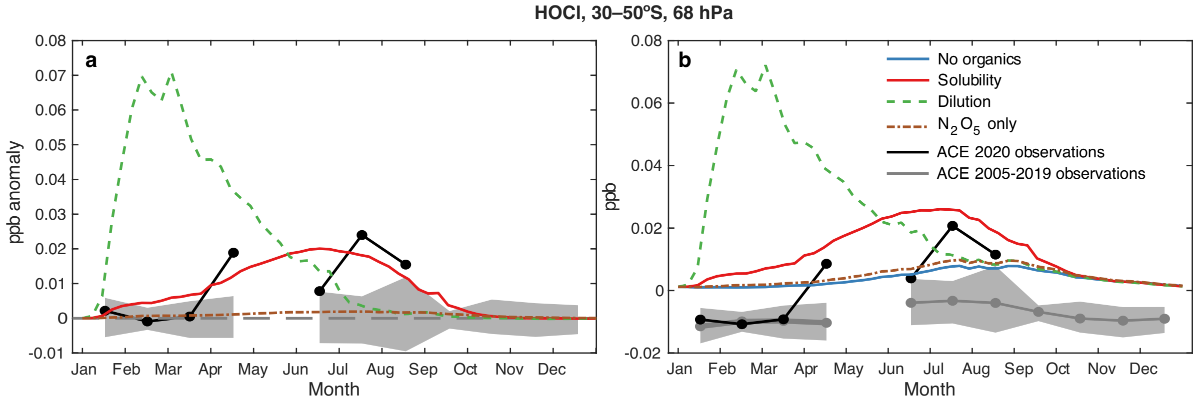
Reprints and permissions information is available at http://www.nature.com/reprints.



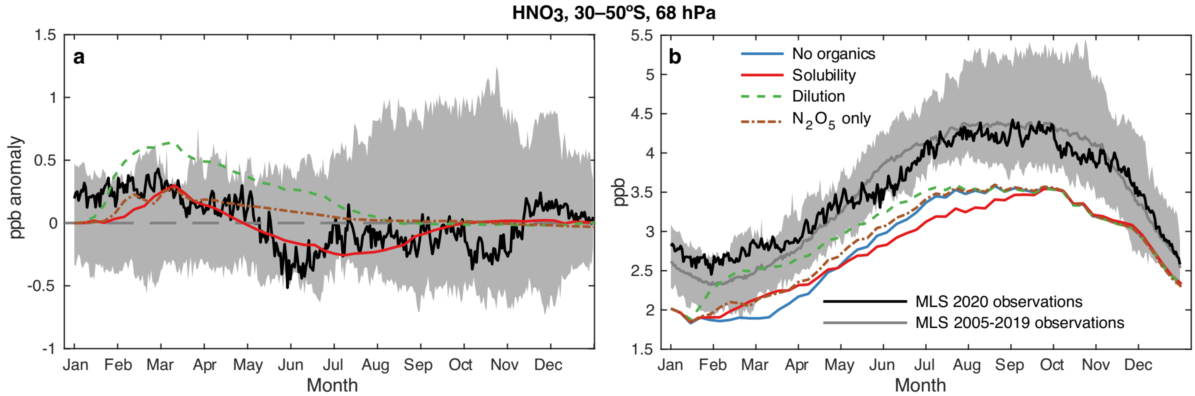
Extended Data Fig. 1 | Modelled and observed aerosol extinction at 18.5 km. The time evolution of aerosol extinction (km^{-1}) is shown at 18.5 km for Ozone Mapping and Profiler Suite (OMPS) observations (for 745 nm, **a**) and in the model (for 675 nm, **b**) in 2020.



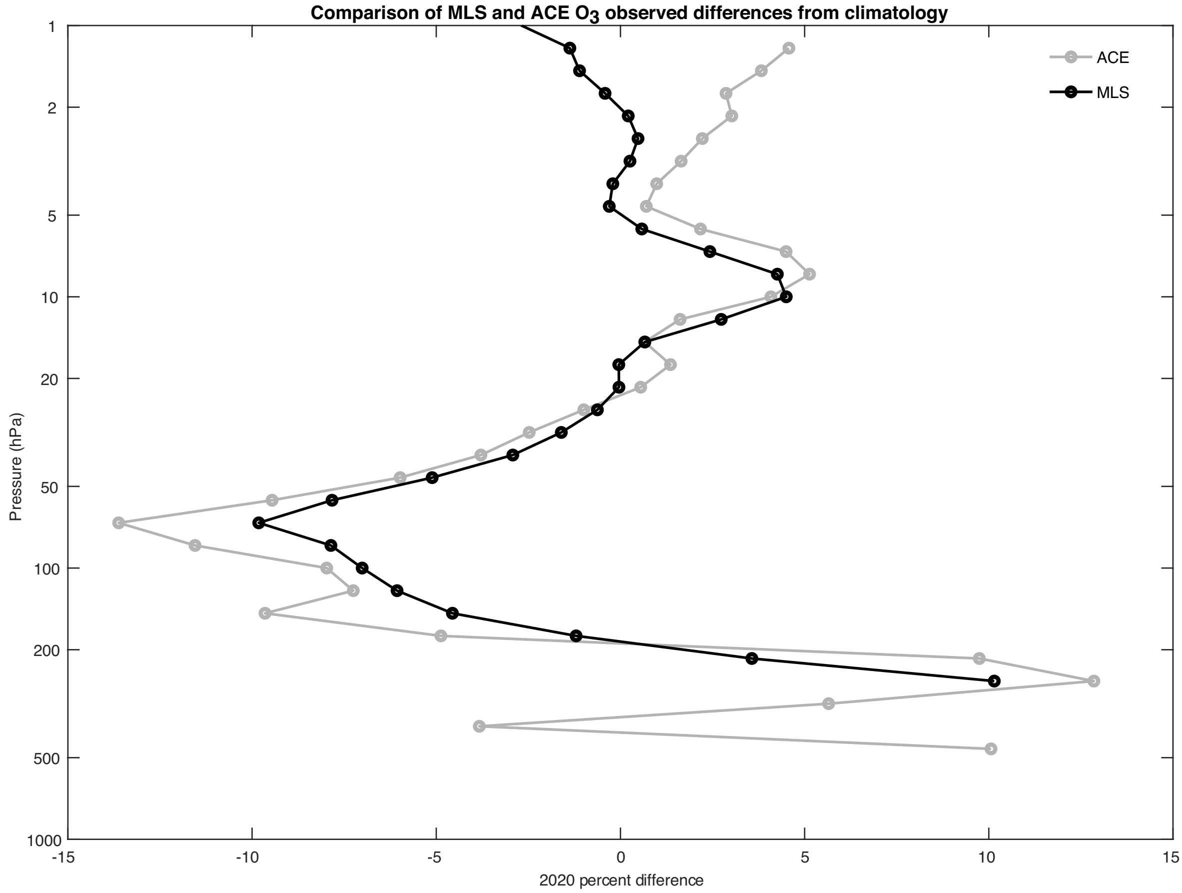
Extended Data Fig. 2 | Observed and modelled 2020 absolute abundances for chemical species from 30–50° S at 68 hPa. Grey shaded regions show the ranges of 24-h averaged satellite data from the climatologies of satellite observations (in mixing ratio units) before 2020 (daily O_3, HCl and ClO from MLS and monthly $ClONO_2$ from ACE) and the grey line shows their averages, whereas black lines show the observed values for 2020. Other coloured lines show model-calculated abundances for the no organics control run (blue line) and for three model test cases: including only N_2O_5 hydrolysis on the aerosols (dashed brown line), considering the added organic material as a dilution factor (green dashed line) and considering the adopted solubility of HCl in organic acid particles (red line). Corresponding anomalies are shown in Fig. 2.



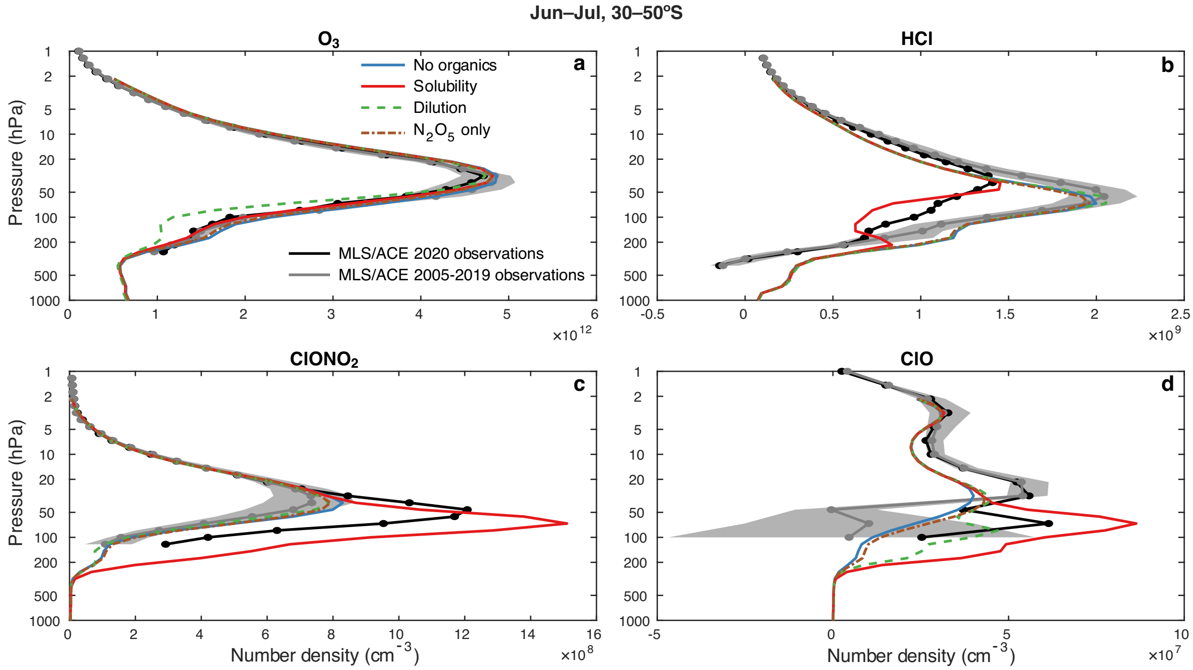
Extended Data Fig. 3 | Observed and modelled monthly averaged anomalies (a) and mixing ratios (b) for HOCl (from ACE) for 30–50° S at 68 hPa. Grey shaded regions show the ranges of 24-h averaged satellite data from the climatology before 2020, whereas black lines show the observed values for 2020. Other coloured lines show calculated values for 2020 for the no organics control run (blue line) and for three model test cases: including only N_2O_5 hydrolysis on the aerosols (dashed brown line), considering the added organic material as a dilution factor (green dashed line) and considering the adopted solubility of HCl in organic acid particles (red line).



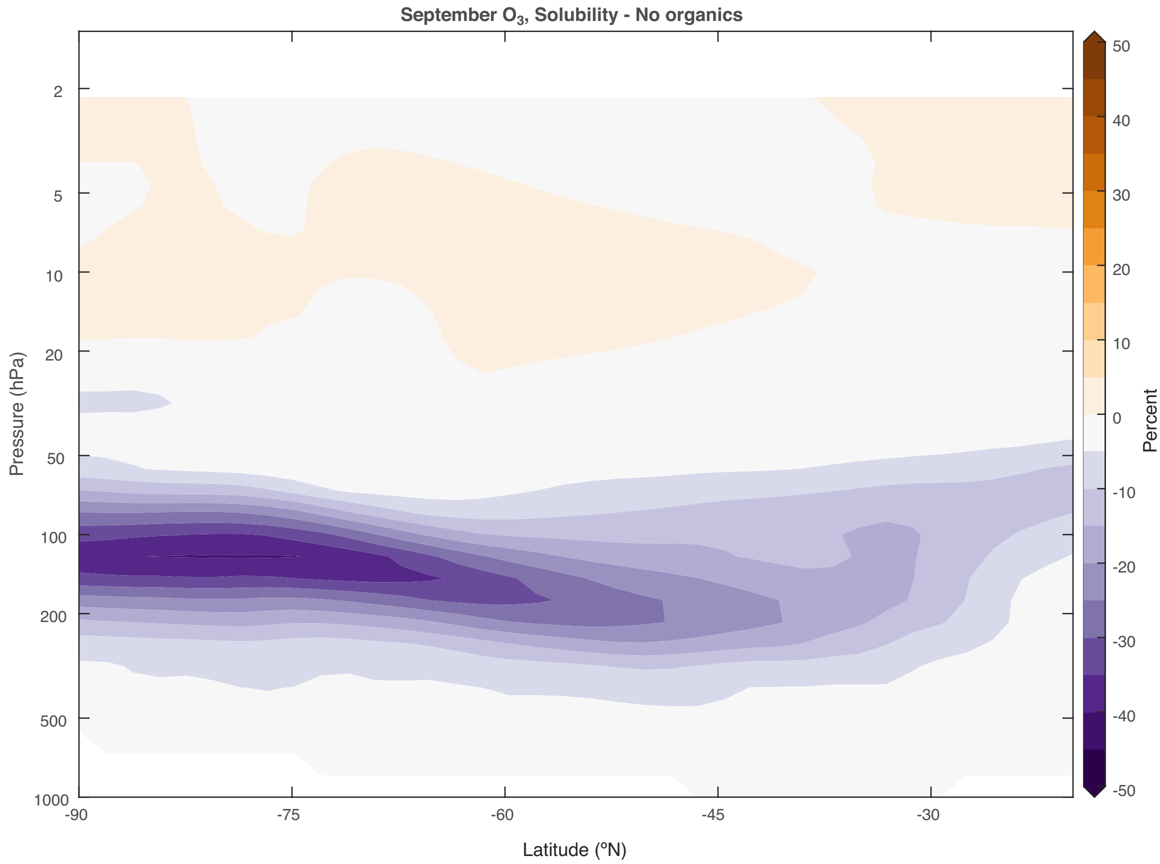
Extended Data Fig. 4 | Observed and modelled anomalies (a) and mixing ratios (b) for HNO_3 (from MLS) for 30–50° S at 68 hPa. Grey shaded regions show the ranges of 24-h daily averaged satellite data from the climatology before 2020, whereas black lines show the observed values for 2020. Other coloured lines show calculated values for 2020 for the no organics control run (blue line) and for three model test cases: including only N_2O_5 hydrolysis on the aerosols (brown dashed line), considering the added organic material as a dilution factor (dashed green line) and considering the adopted solubility of HCl in organic acid particles (red line).



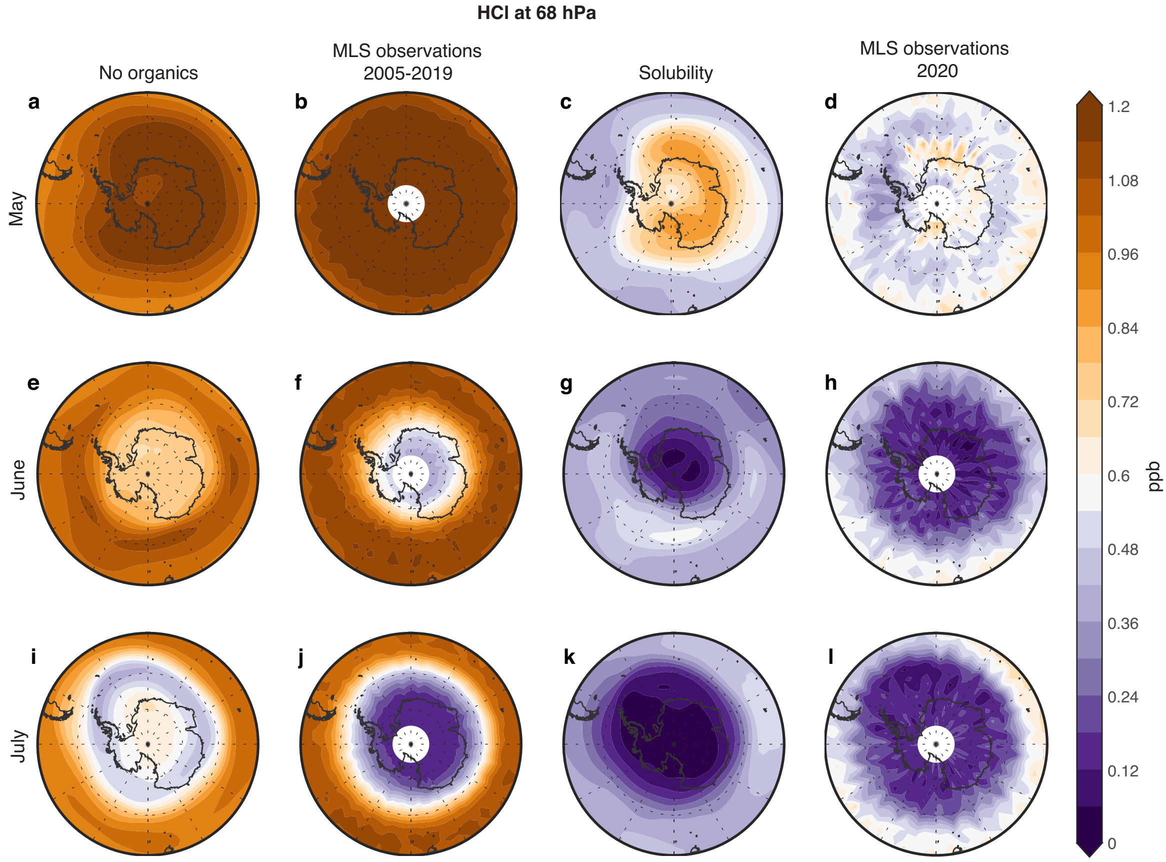
Extended Data Fig. 5 | Percent ozone anomalies for 30–50° S on coincident days of measurement for ACE and MLS during June–July 2020. Data for each satellite have been normalized by their respective climatologies. Note that there are differences in spatial and temporal sampling between the two instruments. Black line shows MLS data while grey line shows ACE data interpolated onto the MLS pressure grid.



Extended Data Fig. 6 | Observed and modelled vertical profile absolute abundances for chemical species from 30–50° S in June–July of 2020. Grey shaded regions show the ranges of 24-h averaged satellite anomalies (in number density units) in years before 2020 (daily O_3 and ClO from MLS and HCl and $ClONO_2$ from ACE) and the grey line shows their averages, whereas black lines show observed abundances for 2020. Other coloured lines show calculated values for 2020 for the no organics control run (blue line) and for three model test cases: including only N_2O_5 hydrolysis on the aerosols (brown dashed line), considering the added organic material as a dilution factor (green dashed line) and considering the adopted solubility of HCl in organic acid particles (red line). Corresponding anomalies are shown in Fig. 3.



Extended Data Fig. 7 | Distribution of calculated ozone loss in September 2020. Percentage change in model-calculated ozone as a function of latitude and height for the oxidized organics solubility model case is shown, as compared with the no organics control run.



Extended Data Fig. 8 | Contour maps of monthly mean HCl abundances (ppbv) at 68 hPa for observations and models. The modelled no organics control case is shown (left column), along with MLS-measured climatological average from 2005–2019 (second from left), modelled oxidized organics solubility case (second from right) and MLS measurements for 2020 (right).

Article

The evolution of the marine carbonate factory

https://doi.org/10.1038/s41586-022-05654-5

Received: 17 May 2022

Accepted: 13 December 2022

Published online: 22 February 2023

Check for updates

Jiuyuan Wang[1]✉, Lidya G. Tarhan[1]✉, Andrew D. Jacobson[2], Amanda M. Oehlert[3] & Noah J. Planavsky[1]

Calcium carbonate formation is the primary pathway by which carbon is returned from the ocean–atmosphere system to the solid Earth[1,2]. The removal of dissolved inorganic carbon from seawater by precipitation of carbonate minerals—the marine carbonate factory—plays a critical role in shaping marine biogeochemical cycling[1,2]. A paucity of empirical constraints has led to widely divergent views on how the marine carbonate factory has changed over time[3–5]. Here we use geochemical insights from stable strontium isotopes to provide a new perspective on the evolution of the marine carbonate factory and carbonate mineral saturation states. Although the production of carbonates in the surface ocean and in shallow seafloor settings have been widely considered the predominant carbonate sinks for most of the history of the Earth[6], we propose that alternative processes—such as porewater production of authigenic carbonates—may have represented a major carbonate sink throughout the Precambrian. Our results also suggest that the rise of the skeletal carbonate factory decreased seawater carbonate saturation states.

The modern marine carbonate factory (that is, the settings, mineralogies and pathways through which carbonates are precipitated from ocean waters[7,8]) is dominated by the formation of biogenic carbonates, including planktonic carbonate tests in the open ocean and skeletal (algal and animal) carbonates in shallow-water environments[6]. However, organismal regulation (for example, biologically controlled precipitation) of carbonates has only been present for a small portion of the history of the Earth[9]. Precambrian carbonates (around 3.5 to 0.54 billion years old) are thought to have been buried largely as inorganic or microbially mediated seafloor precipitates in shallow marine environments[10,11]. Debates surround whether the rise of various pathways of biomineralization caused substantial changes in ocean chemistry and how the locus of major marine carbonate burial sinks may have shifted through the history of the Earth[4]. Although deep-sea carbonate burial is thought to have risen to dominance relatively recently, during the Mesozoic[4], this too has been a subject of contention; previous studies have postulated the existence of a deep-sea carbonate sink before the emergence of planktonic calcifiers (for example, refs. [3,12]). Some studies have suggested that carbonate mineral saturation state (Ω_{carb}), the dominant factor controlling the sequestration of dissolved inorganic carbon into mineral form, has substantially declined through the history of the Earth[1,13], whereas others have argued that the oceans—particularly the surface ocean—have not experienced notable secular changes in carbonate saturation state[5].

Here we present a new approach to tracking the long-term evolution of the marine carbonate factory, based on the record of carbonate stable strontium isotope ratios ($\delta^{88/86}$Sr, for which $\delta^{88/86}\mathrm{Sr} = ((^{88}\mathrm{Sr}/^{86}\mathrm{Sr})_{sample}/(^{88}\mathrm{Sr}/^{86}\mathrm{Sr})_{NIST\text{-}SRM987} - 1) \times 1{,}000$). Recent analytical advances in measuring $\delta^{88/86}$Sr values present a unique opportunity to constrain the evolution of the marine carbonate factory[14–16]. Carbonate $\delta^{88/86}$Sr values are shaped foremost by the relatively large isotope fractionation factor ($\Delta^{88/86}\mathrm{Sr}_{carb\text{-}sw}$) that accompanies carbonate mineral precipitation[14,15,17]. Carbonate precipitation preferentially incorporates lighter Sr isotopes, resulting in low $\delta^{88/86}$Sr values in carbonate minerals and high $\delta^{88/86}$Sr values in remaining fluids[17]. For both biogenic and inorganic (non-skeletal) carbonate minerals, precipitation kinetics strongly regulate Sr isotope fractionation, for which higher Ω_{carb} leads to higher precipitation rates, a greater fractionation (larger-magnitude $\Delta^{88/86}\mathrm{Sr}_{carb\text{-}sw}$) and lower carbonate $\delta^{88/86}$Sr values[17–19]. This kinetic control similarly characterizes different carbonate minerals (that is, aragonite versus calcite), with no marked difference in fractionation during precipitation[18]. Although temperature may affect the physiology of calcifying organisms[19,20] (for instance, growth rate), the dependence of the Sr isotope fractionation on temperature is muted and not currently resolvable[18]. Therefore carbonate $\delta^{88/86}$Sr values can be used to track rates of carbonate mineral precipitation, providing an independent and novel proxy for the evolution of marine carbonate chemistry. Furthermore, when subjected to careful petrographic and geochemical screening, carbonate $\delta^{88/86}$Sr signatures can remain intact even through early and mesodiagenesis (see Methods).

To investigate changes in the marine carbonate factory over a broad period of Earth's early history, we analysed the stable and radiogenic Sr isotope compositions ($\delta^{88/86}$Sr and $^{87}\mathrm{Sr}/^{86}\mathrm{Sr}$) of 115 samples from 29 Precambrian successions ranging from 2.8 billion years to 645 million years in age. These analyses used both double-spike thermal ionization mass spectrometry (TIMS) and Zr-doping multicollector inductively coupled plasma mass spectrometry (MC-ICP-MS). All samples were selected from carbonates reposited in the Yale Peabody Museum's Division of Mineralogy & Meteoritics and which have undergone petrographic and geochemical screening for this and previous studies[21].

[1]Department of Earth and Planetary Sciences, Yale University, New Haven, CT, USA. [2]Department of Earth and Planetary Sciences, Northwestern University, Evanston, IL, USA. [3]Rosenstiel School of Marine, Atmospheric, and Earth Science, University of Miami, Miami, FL, USA. ✉e-mail: jiuyuan.wang@yale.edu; lidya.tarhan@yale.edu

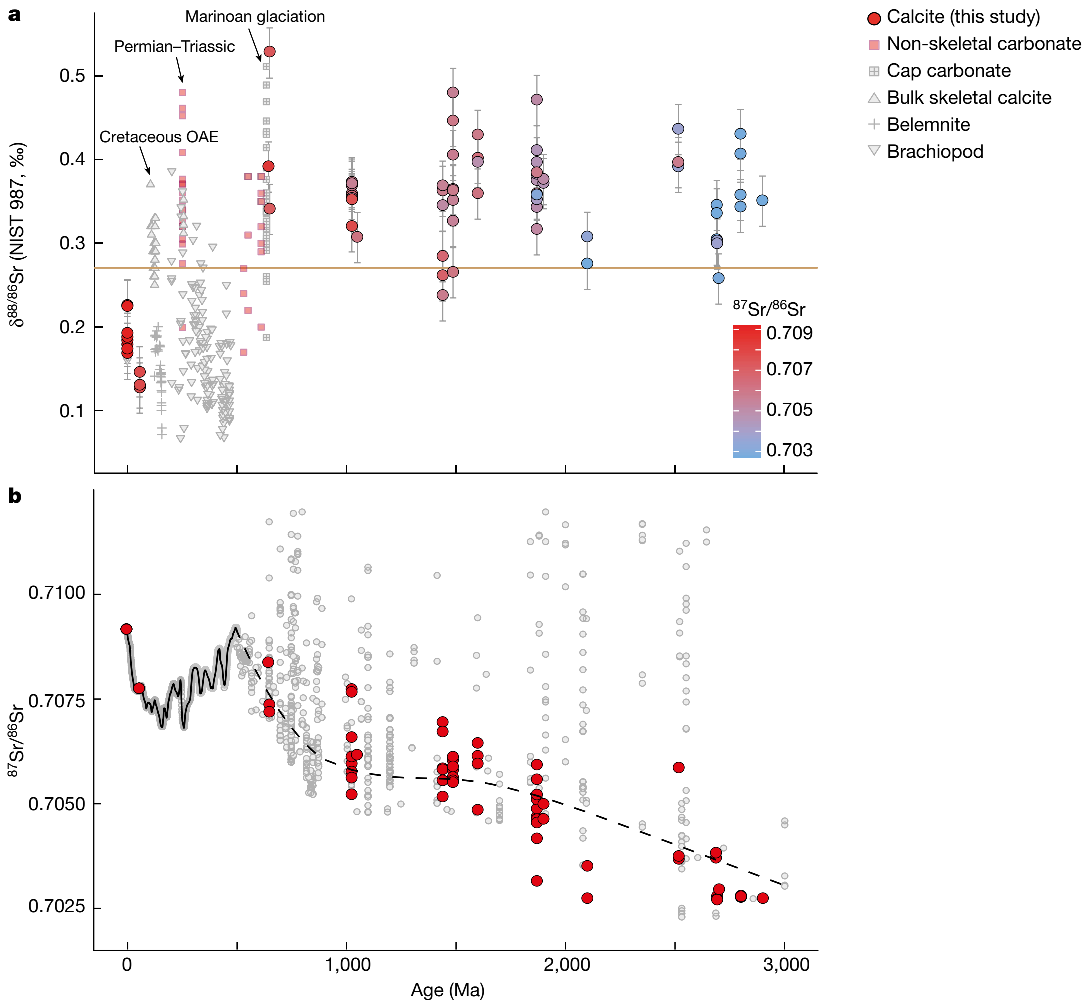


Fig. 1 | Strontium isotope records in marine carbonates through the history of the Earth. a, Summary of $\delta^{88/86}$Sr values measured in skeletal and non-skeletal calcites. New data from this study are shown by the circles ($n = 83$), with the colour scale corresponding to radiogenic Sr isotope ratios (^{87}Sr/^{86}Sr). Error bars represent the long-term external reproducibility of $\delta^{88/86}$Sr ($2\sigma_{SD} = \pm0.03$‰, $n = 273$). The gold line illustrates the $\delta^{88/86}$Sr value of bulk silicate Earth (0.27‰)[27]. Other symbols represent published data from other studies ($n = 299$; Methods). **b**, Marine carbonate ^{87}Sr/^{86}Sr values. New data from this study are shown as red circles. Grey circles represent previously published Precambrian carbonate ^{87}Sr/^{86}Sr records ($n = 1,494$)[23]. The dashed curve denotes the LOESS fit of the lowest 10% of Precambrian ^{87}Sr/^{86}Sr ratios[23] and the solid curve denotes the LOESS fit of the Phanerozoic ^{87}Sr/^{86}Sr record[23]. OAE, oceanic anoxic event.

Owing to subduction of most pre-Mesozoic deep-sea deposits, this sample set (like much of the preserved Precambrian sedimentary record) is dominated by carbonates formed in shallow marine (for example, platform, ramp and reef) settings. During selection of these Precambrian samples, we targeted only non-skeletal shallow marine carbonates and intentionally avoided intervals previously linked to either short-term carbon-cycle perturbations (for example, associated with carbon isotope excursions) or petrographic or geochemical evidence of diagenetic alteration[21] (see Methods). Exogenous Sr inputs from siliciclastic materials and late-stage fluids, such as brine, groundwater and meteoric fluids, tend to increase marine carbonate radiogenic Sr ratios (^{87}Sr/^{86}Sr; for example, ref. [22]). To ensure that our selected samples provide a reliable seawater Sr archive, the measured ^{87}Sr/^{86}Sr ratio of each sample was also compared with ^{87}Sr/^{86}Sr records widely considered to record contemporaneous seawater[23]. Among our petrographically and geochemically screened samples, we regard calcites with ^{87}Sr/^{86}Sr ratios closely aligned with contemporaneous seawater ^{87}Sr/^{86}Sr ratios[23] to be the best preserved and most representative. To better facilitate comparison with recent marine sediments, we also analysed 24 Palaeogene and modern carbonate samples spanning marginal marine to deep-sea settings. To supplement our Precambrian and recent datasets, we further compiled published radiogenic and stable Sr isotope data for both skeletal and non-skeletal Phanerozoic and Ediacaran calcites ($n = 287$). All datasets include microbially mediated (for example, stromatolitic) as well as 'abiotic' non-skeletal carbonates.

The most notable first-order feature observed in our composite $\delta^{88/86}$Sr record is the significant distinction between Precambrian and Phanerozoic calcite $\delta^{88/86}$Sr values (Fig. 1; t-test $P = 7.24 \times 10^{-21}$). The average $\delta^{88/86}$Sr value for the Precambrian (and pre-Ediacaran) portion of our record (mean = 0.36‰; 2,800 to 645 million years ago, or Ma) is 0.20‰ higher than Phanerozoic values (mean = 0.16‰; 540 Ma to present), with a sharp decrease around the Ediacaran (635 to 540 Ma). The highest values in the Phanerozoic are associated with two massive carbon-cycle perturbations (Fig. 2), namely the end-Permian mass extinction (mean = 0.35‰; 252 Ma)[24] and the Early Cretaceous Oceanic Anoxic Event 1a (mean = 0.29‰; 120 Ma)[15], and are not representative of background Phanerozoic conditions. The aftermath of another notable climatic perturbation, the Marinoan Snowball Earth (which terminated at 635 Ma), is also characterized by elevated $\delta^{88/86}$Sr values (mean = 0.35‰)[25]. These events have each been linked to changes in carbonate saturation state[26], but they are also notable examples of non-steady-state perturbations.

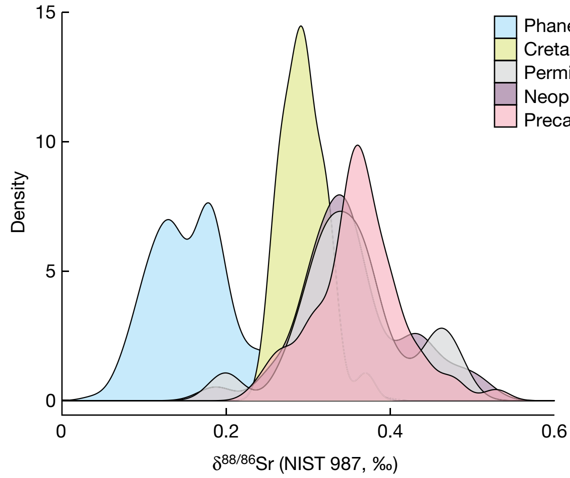


Fig. 2 | Density plots of $\delta^{88/86}$Sr records in different time intervals. Red represents Precambrian calcites (this study); blue represents Phanerozoic calcites from previously published studies (Methods); purple represents Neoproterozoic Marinoan cap carbonates ($n = 36$)[25]; grey represents skeletal and non-skeletal calcites spanning the uppermost Permian and lowest Triassic ($n = 34$)[24]; green represents skeletal calcites formed during the Early Cretaceous Oceanic Anoxic Event 1a ($n = 32$)[15]. OAE, oceanic anoxic event.

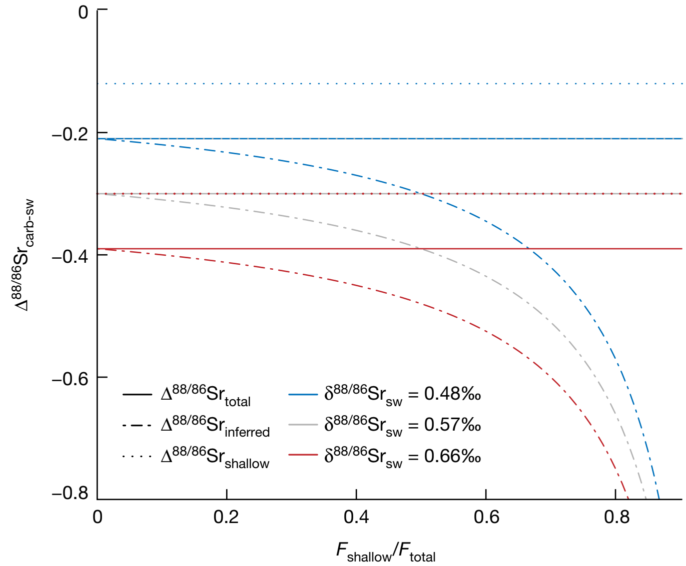


Fig. 3 | Mass balance regulating the impact of seawater saturation state and variable partitioning between a shallow marine (platform, ramp and reef) carbonate burial sink and an inferred isotopically depleted sink on carbonate Sr isotope fractionation. The solid line represents the explicit Sr isotope fractionation of all carbonate sinks ($\Delta^{88/86}Sr_{total}$); the dotted and dashed lines denote the Sr isotope fractionation of the shallow marine ($\Delta^{88/86}Sr_{shallow}$) and inferred isotopically depleted ($\Delta^{88/86}Sr_{inferred}$) carbonate sinks, respectively. Colours indicate different choices of $\delta^{88/86}Sr_{sw}$: blue (0.48‰) reflects a total carbonate fractionation comparable with that governing precipitation of modern marine carbonates ($\Delta^{88/86}Sr_{total} = -0.21$‰), grey (0.57‰) reflects a modern-like fractionation for shallow marine carbonates ($\Delta^{88/86}Sr_{shallow} = -0.21$‰) and red (0.66‰) represents a hypothetical larger fractionation associated with precipitation under elevated carbonate saturation state ($\Delta^{88/86}Sr_{shallow} = -0.30$‰).

Critically, the observed temporal shift between Precambrian and Phanerozoic marine carbonate $\delta^{88/86}$Sr values cannot be readily linked to diagenetic alteration. Analysed Precambrian dolomites and selected calcites with $^{87}Sr/^{86}Sr$ ratios higher than estimated seawater values[23]—and thus most parsimoniously interpreted as altered—are, on average, characterized by $\delta^{88/86}$Sr values notably lower (mean = 0.22‰; Extended Data Fig. 1a) than those of well-preserved Precambrian calcites. The disparity in $\delta^{88/86}$Sr values between dolomite and well-preserved calcite is even more evident following bootstrap resampling (Extended Data Fig. 2b; *t*-test $P = 6.51 \times 10^{-34}$). Late-stage alteration encompasses a range of processes that would impart different $^{87}Sr/^{86}Sr$ and $\delta^{88/86}$Sr signatures. Nonetheless, our data suggest that alteration of Precambrian carbonates leads to lower carbonate $\delta^{88/86}$Sr values (Methods). The sustained high $\delta^{88/86}$Sr values of the Precambrian (that is, higher than the average value of bulk silicate Earth; $\delta^{88/86}Sr_{BSE} = 0.27$‰ in ref. [27]) have two direct implications: (1) the $\delta^{88/86}$Sr of the Precambrian oceans must have been high owing to preferential uptake of lighter Sr during carbonate precipitation; (2) these isotopically enriched non-skeletal carbonates could not alone have maintained Sr isotope mass balance (Supplementary Discussion A). The $\delta^{88/86}$Sr value of the integrated input flux should, over geologic (10^6-year) timescales, approximate $\delta^{88/86}Sr_{BSE}$ (0.27‰)[27]. This apparent discrepancy suggests (1) that a substantial sink of lighter Sr isotopes must have existed beyond deposition of carbonates in the shallow marine environments we targeted (carbonate platforms, reefs and ramps) and (2) that the Sr isotope fractionation, seawater $\delta^{88/86}$Sr value or the relative abundance of these sinks (or some combination thereof) must have varied in Earth's past. Specifically, some marine carbonates must have precipitated under larger-magnitude $\Delta^{88/86}Sr_{carb\text{-}sw}$ values, linked to more rapid precipitation rates (and higher Ω_{carb} at the site of precipitation[17,19,20]).

Mass-balance constraints can be used to link our new stable Sr isotope record to changes in marine carbonate chemistry (Fig. 3; Supplementary Discussion A). We cannot precisely constrain ancient seawater $\delta^{88/86}$Sr values. Nonetheless, we can explore possible global Sr isotope mass balances by making simple assumptions about seawater–carbonate isotope fractionations ($\Delta^{88/86}Sr_{carb\text{-}sw}$). An endmember case would be to assume a total (all sinks) seawater–carbonate fractionation comparable with that of the modern oceans ($\Delta^{88/86}Sr_{total}$ of −0.21‰ (ref. [14])) and that the Sr input value was the same as the bulk silicate Earth ($\delta^{88/86}Sr_{input} = 0.27$‰ (ref. [27])). Using the average value of our measured Precambrian shallow marine carbonates (0.36‰), this translates into a shallow marine carbonate fractionation of −0.12‰. Alternatively, if shallow marine carbonates showed a similar seawater fractionation as in the modern oceans ($\Delta^{88/86}Sr_{shallow}$ of −0.21‰ (ref. [14])), $\Delta^{88/86}Sr_{total}$ would be −0.30‰. The heaviest pre-Cryogenian carbonates in our record (0.48‰) would, with a modern $\Delta^{88/86}Sr_{shallow}$ value, be linked to a $\Delta^{88/86}Sr_{total}$ of −0.42‰. This requires that carbonates not represented by our shallow marine record have even larger-magnitude $\Delta^{88/86}Sr_{shallow}$ values. And if Precambrian global marine carbonate saturation states were, as predicted[13], higher than in the modern ocean (driving larger-magnitude $\Delta^{88/86}Sr_{shallow}$ values), even larger-magnitude $\Delta^{88/86}Sr_{total}$ values are required. Mass balance demonstrates that our $\delta^{88/86}$Sr record requires substantial carbonate burial outside typical shallow-water carbonate factories (Fig. 3; Supplementary Discussion A), provoking the question: what environmental settings or phases are the most probable candidates for this 'missing' carbonate burial sink?

Microbial carbonates are common components throughout much of the Precambrian sedimentary record (for example, ref. [11]) and anaerobic microbial activity can locally increase Ω_{carb} values. Microbially mediated carbonate (for example, stromatolite) precipitation can occur under Ω_{carb} values elevated with respect to ambient seawater[28]. However, the high $\delta^{88/86}$Sr values we observe in Precambrian carbonates cannot simply be attributed to a physiological influence on precipitation or a shift in the relative abundance of skeletal carbonate; microbial carbonates, non-microbial carbonates and skeletal carbonates are characterized by statistically indistinguishable $\delta^{88/86}$Sr values (Extended Data Fig. 5; Supplementary Discussion B). Other potential isotopically depleted mineral phases—such as clays—are unlikely to

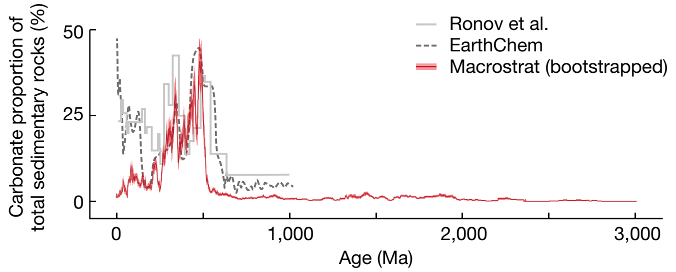


Fig. 4 | Carbonate relative abundance in preserved sedimentary rocks through time. The grey curve represents carbonate relative abundance estimates from all present-day continents (excluding Antarctica) by Ronov et al.[34]; the dotted curve illustrates the estimate of global carbonate relative abundance inferred by CaO content using the EarthChem database, which also includes marine drill core data; the red curve represents measurements of carbonate relative abundance within North American sedimentary rock columns ($n = 949$) using the Macrostrat database; the red-shaded range represents bootstrap resampling ($n = 10{,}000$) of the Macrostrat estimates for every 1-million-year interval.

have had a substantial impact on the global Sr isotope mass balance (Supplementary Discussion C). Therefore we propose that this 'missing' reservoir (or potentially several reservoirs) of isotopically light Sr was stored in undercharacterized phases, such as authigenic carbonates precipitated in muddy seafloor sediments from intergranular fluids (that is, porewaters). This sink could have formed in a range of seafloor settings, both shallow and deep marine.

Authigenic carbonates have previously been proposed to represent an important Precambrian carbon sink[8]. We cannot, from our current dataset, directly isolate which potential authigenic sinks may have been most prominent. However, environments characterized by higher rates of organic carbon delivery and anaerobic remineralization pathways such as iron reduction (for example, muddy continental margins[29] and, in the Precambrian, potentially also the deep sea) may, particularly under the low-oxygen and ferruginous bottom-water conditions characteristic of much of the Precambrian, have been a particularly important locus of authigenic carbonate precipitation[5,8]. Moreover, limited bioturbation in the Precambrian would have dampened diffusive loss of porewater alkalinity (for example, generated by means of iron reduction) and facilitated greater oversaturation within seafloor sediments than in the overlying water column[30]. Higher baseline marine carbonate saturation states (high Ω_{carb}) could have also fostered extensive early diagenetic carbonate formation[31]. Although current knowledge of stable Sr isotope systematics is still evolving and our records do not preclude the influence of other factors (Supplementary Discussions C and D), a substantial authigenic carbonate burial flux characterized by low $\delta^{88/86}Sr_{carb}$ values under high Ω_{carb} and formed from upper-seafloor porewaters (that is, under relatively open-system conditions) provides a simple means to satisfy the Sr isotope mass balance. Therefore our new record provides empirical support for a markedly different view of the Precambrian global carbon cycle[8]. Intriguingly, similar patterns have not been observed in the long-term Ca isotope ($\delta^{44/40}Ca$) record[32], suggesting a decoupling—potentially during early diagenesis—between these two isotope systems (Supplementary Discussion E).

The transition from high to lower $\delta^{88/86}Sr_{carb}$ values coincides with the innovation of widespread skeletal biomineralization in the Ediacaran and Palaeozoic. We suggest that shallow marine seafloor settings, in which most early calcifying organisms thrived, became the predominant locus of carbonate production during this interval. This idea is consistent with existing records of the relative proportion of all preserved sedimentary rocks represented by carbonates, which suggest notable increases around the Ediacaran and Palaeozoic and subsequent inflections across the Mesozoic–Cenozoic transition (Fig. 4). This latter transition, coinciding with the Mesozoic radiation of calcareous plankton, may reflect the corresponding growth of the pelagic deep-sea carbonate sink and the concomitant decline of shallow marine sinks, because these compilations are dominated by cratonic (continental and shallow marine) deposits (with the exception of the EarthChem dataset, which includes marine drill core data). Although all current sedimentary databases offer incomplete snapshots of the sedimentary record, the markedly low relative abundance of carbonate in Precambrian sedimentary compilations may indicate that, during the Precambrian, the canonical shallow marine carbonate factory represented a smaller sink than historically imagined.

The expansion of biologically controlled pathways of biomineralization, in which organisms exert some amount of control over intercellular or intracellular carbonate chemistry and precipitation rates (for example, ref. [33]), would have fundamentally transformed the carbonate factory and may have led to a Palaeozoic acme in the precipitation and preservation of shallow marine carbonates (Fig. 4). Our results bolster the case that this transition to widespread skeletal biomineralization would have led to a drop in baseline seawater Ω_{carb} and alkalinity and, eventually, with the Mesozoic rise of calcareous plankton, the stabilization of Earth's marine carbon cycle[4].

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-05654-5.

1. Ridgwell, A. & Zeebe, R. E. The role of the global carbonate cycle in the regulation and evolution of the Earth system. *Earth Planet. Sci. Lett.* **234**, 299–315 (2005).
2. Isson, T. T. et al. Evolution of the global carbon cycle and climate regulation on Earth. *Global Biogeochem. Cycles* **34**, 1–28 (2020).
3. Wilkinson, B. H. & Walker, J. C. G. Phanerozoic cycling of sedimentary carbonate. *Am. J. Sci.* **289**, 525–548 (1989).
4. Ridgwell, A. A Mid Mesozoic Revolution in the regulation of ocean chemistry. *Mar. Geol.* **217**, 339–357 (2005).
5. Higgins, J. A., Fischer, W. W. & Schrag, D. P. Oxygenation of the ocean and sediments: consequences for the seafloor carbonate factory. *Earth Planet. Sci. Lett.* **284**, 25–33 (2009).
6. James, N. P. & Jones, B. *Origin of Carbonate Sedimentary Rocks* (Wiley, 2015).
7. Schlager, W. Sedimentation rates and growth potential of tropical, cool-water and mud-mound carbonate systems. *Geol. Soc. Spec. Publ.* **178**, 217–227 (2000).
8. Schrag, D. P., Higgins, J. A., Macdonald, F. A. & Johnston, D. T. Authigenic carbonate and the history of the global carbon cycle. *Science* **339**, 540–543 (2013).
9. Gilbert, P., Bergmann, K. & Knoll, A. H. Biomineralization: integrating mechanism and evolutionary history. *Sci. Adv.* **8**, eabl9653 (2021).
10. Grotzinger, J. P. & James, N. P. in *Carbonate Sedimentation and Diagenesis in the Evolving Precambrian World* 3–20 (SEPM Society for Sedimentary Geology, 2000).
11. Cantine, M. D., Knol, A. H. & Bergmann, K. D. Carbonates before skeletons: a database approach. *Earth-Sci. Rev.* **201**, 103065 (2020).
12. Simonson, B. M., Schubel, K. A. & Hassler, S. W. Carbonate sedimentology of the early Precambrian Hamersley Group of Western Australia. *Precambrian Res.* **60**, 287–335 (1993).
13. Grotzinger, J. P. Geochemical model for Proterozoic stromatolite decline. *Am. J. Sci.* **290**, 80–103 (1990).
14. Vollstaedt, H. et al. The Phanerozoic $\delta^{88/86}Sr$ record of seawater: new constraints on past changes in oceanic carbonate fluxes. *Geochim. Cosmochim. Acta* **128**, 249–265 (2014).
15. Wang, J., Jacobson, A. D., Sageman, B. B. & Hurtgen, M. T. Stable Ca and Sr isotopes support volcanically triggered biocalcification crisis during Oceanic Anoxic Event 1a. *Geology* **49**, 515–519 (2021).
16. Paytan, A. et al. A 35-million-year record of seawater stable Sr isotopes reveals a fluctuating global carbon cycle. *Science* **371**, 1346–1350 (2021).
17. Böhm, F. et al. Strontium isotope fractionation of planktic foraminifera and inorganic calcite. *Geochim. Cosmochim. Acta* **93**, 300–314 (2012).
18. AlKhatib, M. & Eisenhauer, A. Calcium and strontium isotope fractionation in aqueous solutions as a function of temperature and reaction rate; I. Calcite. *Geochim. Cosmochim. Acta* **209**, 296–319 (2017).
19. Müller, M. N., Krabbenhöft, A., Vollstaedt, H., Brandini, F. P. & Eisenhauer, A. Stable isotope fractionation of strontium in coccolithophore calcite: influence of temperature and carbonate chemistry. *Geobiology* **16**, 297–306 (2018).
20. Stevenson, E. I. et al. Controls on stable strontium isotope fractionation in coccolithophores with implications for the marine Sr cycle. *Geochim. Cosmochim. Acta* **128**, 225–235 (2014).
21. Kalderon-Asael, B. et al. A lithium-isotope perspective on the evolution of carbon and silicon cycles. *Nature* **595**, 394–398 (2021).

22. Banner, J. L. Application of the isotope and trace element geochemistry of strontium to studies of diagenesis in carbonate systems. *Sedimentology* **42**, 805–824 (1995).
23. Shields, G. & Veizer, J. Precambrian marine carbonate isotope database: version 1.1. *Geochem. Geophys. Geosyst.* 3, 10.1029/2001GC000266 (2002).
24. Wang, J. et al. Coupled $\delta^{44/40}$Ca, $\delta^{88/86}$Sr, and ^{87}Sr/^{86}Sr geochemistry across the end-Permian mass extinction event. *Geochim. Cosmochim. Acta* **262**, 143–165 (2019).
25. Wang, J., Jacobson, A. D., Sageman, B. B. & Hurtgen, M. T. $\delta^{44/40}$Ca-$\delta^{88/86}$Sr multi-proxy constrains primary origin of Marinoan cap carbonates. Preprint at https://arxiv.org/abs/2204.02563 (2022).
26. Kump, L., Bralower, T. & Ridgwell, A. Ocean acidification in deep time. *Oceanography* **22**, 94–107 (2009).
27. Moynier, F., Agranier, A., Hezel, D. C. & Bouvier, A. Sr stable isotope composition of Earth, the Moon, Mars, Vesta and meteorites. *Earth Planet. Sci. Lett.* **300**, 359–366 (2010).
28. Dupraz, C. et al. Processes of carbonate precipitation in modern microbial mats. *Earth-Sci. Rev.* **96**, 141–162 (2009).
29. Dunne, J. P., Sarmiento, J. L. & Gnanadesikan, A. A synthesis of global particle export from the surface ocean and cycling through the ocean interior and on the seafloor. *Global Biogeochem. Cycles* **21**, GB4006 (2007).
30. Wright, V. P. & Cherns, L. Leaving no stone unturned: the feedback between increased biotic diversity and early diagenesis during the Ordovician. *J. Geol. Soc.* **173**, 241–244 (2016).
31. Laakso, T. A. & Schrag, D. P. The role of authigenic carbonate in Neoproterozoic carbon isotope excursions. *Earth Planet. Sci. Lett.* **549**, 116534 (2020).
32. Blättler, C. L. & Higgins, J. A. Testing Urey's carbonate–silicate cycle using the calcium isotopic composition of sedimentary carbonates. *Earth Planet. Sci. Lett.* **479**, 241–251 (2017).
33. Weiner, S. & Dove, P. M. An overview of biomineralization processes and the problem of the vital effect. *Rev. Mineral. Geochem.* **54**, 1–29 (2003).
34. Ronov, A. B., Khain, V. E., Balukhovsky, A. N. & Seslavinsky, K. B. Quantitative analysis of Phanerozoic sedimentation. *Sediment. Geol.* **25**, 311–325 (1980).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2023

Methods

Materials and sample preparation

All studied Precambrian carbonate samples were selected from samples previously reposited in the Yale Peabody Museum's Division of Mineralogy and Meteoritics. Any intervals previously linked to short-term carbon-cycle perturbations or diagenetic alteration (that is, associated with carbon isotope excursions[35]) were intentionally avoided. Strict petrographic and geochemical screening criteria were applied to select the least-altered carbonate samples (see below for details). A total of 437 samples were petrographically examined, 276 samples were measured for elemental concentrations and 115 samples were selected for Sr isotope analyses. The selected Precambrian samples are either from drill cores or from the least-altered portions of outcrops interpreted to record deposition in shallow marine (platform, ramp and reef) environments (Supplementary Table 1). Previous publications describing the ages, lithologies and associated depositional environments of the samples are listed in Supplementary Table 1. Drill core samples were pre-washed, dried and carefully selected to avoid any fractures or veins. Field samples were taken from well-preserved outcrops that were not characterized by faulting or veins and had not experienced anthropogenic modification (for example, fracking). Hand samples were cut into slabs for making thin sections, sampling and archiving. All thin sections were petrographically screened under optical and, in some cases, cathodoluminescence microscopy by Kalderon-Asael et al.[21]. In general, micritic carbonates, microbialites, ooids and marine cements that showed the highest degree of petrographic preservation and lowest abundance of clay were targeted. The lithological facies of selected samples are noted in Supplementary Table 1. Well-preserved areas that represent the bulk matrix of the carbonate were micro-drilled into powders with a low-speed, stationary micro-drill equipped with a 0.7-mm tungsten-carbide bit. Palaeogene samples were selected from ODP Hole 866A, on the flanks of Resolution Guyot in the mid-Pacific Ocean. Modern carbonate samples were collected from Little Darby Island (Bahamas), Schwab Lake (Florida) and Hamelin Pool, Shark Bay (Australia), and are also reposited in the Yale Peabody Museum's Division of Mineralogy and Meteoritics. Modern deep-sea samples were collected from the core-top repository at Woods Hole Oceanographic Institution. Detailed sample locations and lithologies for Palaeogene and modern carbonate samples are listed in Supplementary Table 2.

For elemental and isotope analyses, we used a modified sequential leaching method[24,36] to extract the most representative portion of the primary carbonate without affecting non-carbonate phases. Briefly, approximately 200 mg of powder was weighed into acid-clean 50-ml centrifuge tubes. To separate surface-adsorbed phases and any exchangeable cations, weighed carbonate samples were first leached with 25 ml of 1 N ammonium acetate (NH_4Ac, buffered to pH = 8.2) for 30 min on a rocker table and centrifuged for 10 min at 3,600 rpm, and the supernatant was then decanted. The residues were washed in 25 ml Milli-Q water (18.2 MΩ-cm at 25 °C) and the supernatant was discarded after centrifuging. The remaining residues were then sequentially reacted with 25 ml of 0.25% and 1% acetic acid; these sequential reactions were performed twice. For each step, mixtures were agitated in an ultrasonic bath for 15 min to ensure a thorough reaction, then centrifuged to decant the supernatant and then washed with Milli-Q water before proceeding to the next step. After these steps, the calcite samples were dissolved in 25 ml of 5% acetic acid and dolomite samples were dissolved in 25 ml of 5% nitric acid (HNO_3). Both calcite and dolomite samples were gently shaken on a rocker table overnight to ensure complete reaction. The mixtures were centrifuged and the supernatants were passed through acid-cleaned 0.20-μm polypropylene syringe filters (Whatman Puradisc), collected in Teflon beakers, dried at 90 °C, then redissolved in 20 ml of 5% HNO_3 and stored as sample solutions for elemental and isotopic analyses. All procedures used TraceMetal Grade reagents.

Analytical methods

Elemental concentrations. Elemental concentrations (Li, Mg, Al, Ca, Ti, Mn, Rb, Sr and Pb) were measured at Yale University using a Thermo Scientific Element XR ICP-MS equipped with a quartz spray chamber. For each sample solution, a 20-μl aliquot was dried at 90 °C, diluted to 4 ml with 5% HNO_3 to minimize matrix effects and then spiked with In to achieve an In concentration of 2 ppb. Total procedural blanks were treated as samples and monitored throughout the analyses. The geostandards BHVO-2 and SBC-1 were repeatedly analysed every 20 samples to assess the accuracy and reproducibility of the method. Results were within ±5% of reported concentrations for both major and trace elements.

Radiogenic and stable Sr isotope ratios. Two different methods were used to measure Sr isotope ratios ($^{87}Sr/^{86}Sr$ and $\delta^{88/86}Sr$): (1) Zr-doping MC-ICP-MS using a Thermo Neptune Plus at the Yale Metal Geochemistry Center and (2) double-spike TIMS using a Thermo TRITON Plus in the Radiogenic Isotope Geochemistry Clean Laboratory at Northwestern University. All samples were analysed out of stratigraphic order. To ensure data reproducibility, 35% of the samples (n = 41) were measured as duplicates with three different strategies, including reanalysis of the same sample solution, redissolution and reanalysis of the same drilled powder and reprocessing and reanalysis of altogether different powders drilled from the same original hand samples. A handful of samples were measured using each method (Zr-doping MC-ICP-MS and double-spike TIMS) to cross-validate accuracy and confirm intermethod differences were within analytical precision. The entire analytical period spanned over two years.

The Zr-doping MC-ICP-MS measurements followed the method described by Wang et al.[37]. Briefly, aliquots of sample solutions containing 500 ng of Sr were weighed into Teflon vials, dried, redissolved in 1 ml 8N HNO_3 and eluted through Bio-Rad columns packed with 1 ml Eichrom Sr-Spec resin to separate Sr from matrix elements. The purified Sr solutions were dried, reacted with H_2O_2 to eliminate potential organics, refluxed with concentrated HNO_3, redried and finally redissolved with 1 ml of 5% HNO_3. Total procedural blanks during the period of study were approximately 150 pg for Sr (n = 31), which is negligible compared with the amounts of sample Sr processed. Sample solutions, as well as solutions of the NBS 987 bracketing standard, were diluted and doped with a Zr solution to achieve concentrations of 100 ng g^{-1} for Sr and 200 ng g^{-1} for Zr. Samples were separated by two adjacent bracketing NBS 987 standards and a 5% HNO_3 rinse solution between each sample and standard. In the MC-ICP-MS, nine masses (83, 84, 85, 86, 87, 88, 90, 91 and 92) were analysed simultaneously to calculate Sr and Zr isotope ratios, as well as to calculate corrections for Kr and Rb isobaric interferences. Sample $^{87}Sr/^{86}Sr$ and $^{88}Sr/^{86}Sr$ ratios were calculated after rinse correction, interference correction, mass discrimination correction and normalization. All $^{88}Sr/^{86}Sr$ ratios are reported after standard bracketing, in δ notation relative to NBS 987, for which $\delta^{88/86}Sr$ (in ‰) = $((^{88}Sr/^{86}Sr)_{smp}/(^{88}Sr/^{86}Sr)_{NBS987} - 1) \times 1,000$. Repeated measurements of NIST 987 yielded a mean $^{87}Sr/^{86}Sr$ ratio of 0.710250 ± 0.000001 ($2\sigma_{SEM}$, n = 215) and a mean $\delta^{88/86}Sr$ value of 0.000 ± 0.002‰ ($2\sigma_{SEM}$, n = 215). Three reference materials were analysed during the period of study. IAPSO seawater yielded $^{87}Sr/^{86}Sr$ = 0.709175 ± 0.000003 ($2\sigma_{SEM}$, n = 30) and $\delta^{88/86}Sr$ = 0.390 ± 0.005‰ (±$2\sigma_{SEM}$, n = 30). BHVO-2 yielded $^{87}Sr/^{86}Sr$ = 0.703496 ± 0.000003 ($2\sigma_{SEM}$, n = 16) and $\delta^{88/86}Sr$ = 0.255 ± 0.008‰ ($2\sigma_{SEM}$, n = 16). BCR-2 yielded 0.705019 ± 0.000004 ($2\sigma_{SEM}$, n = 12) and $\delta^{88/86}Sr$ = 0.285 ± 0.009‰ ($2\sigma_{SEM}$, n = 12). These results agree well with previously published values[27,37–40] and correspond to a long-term external reproducibility ($2\sigma_{SD}$) of ±0.000015 and ±0.03‰ for $^{87}Sr/^{86}Sr$ and $\delta^{88/86}Sr$, respectively.

The double-spike TIMS measurements followed the procedures presented by Wang et al.[24]. For this method, $^{87}Sr/^{86}Sr$ and $^{88}Sr/^{86}Sr$ ratios are measured in separate mass-spectrometric runs. To measure $^{87}Sr/^{86}Sr$

ratios, sample solutions containing approximately 300 ng of Sr were weighed, dried and redissolved in 8M HNO_3, and then eluted through inverted pipette tips containing roughly 1 ml of Eichrom Sr-Spec resin. After drying and redissolution, approximately 150 ng of Sr was loaded onto outgassed single Re filament assemblies, along with 1 µl of a $TaCl_5$ solution. Measurements were performed in dynamic mode with amplifier rotation. The ^{85}Rb beam was monitored to confirm that ^{87}Rb did not isobarically interfere with ^{87}Sr. Instrumental mass fractionation was corrected by normalizing ^{86}Sr/^{88}Sr ratios to 0.1194 using an exponential law. During the period of study, repeated measurements of NBS 987 yielded a mean ^{87}Sr/^{86}Sr ratio of 0.710250 ± 0.000003 ($2\sigma_{SEM}$, n = 9), consistent with the long-term external reproducibility ($2\sigma_{SD}$) of ±0.000010, which is the uncertainty assigned to the TIMS ^{87}Sr/^{86}Sr measurements. To measure ^{88}Sr/^{86}Sr ratios, sample solutions containing 450 ng of Sr were weighed and mixed with an ^{87}Sr–^{84}Sr double spike. After refluxing at 90 °C overnight, the solutions were dried, redissolved in 8N HNO_3 and purified using the same elution procedure described above. After drying and redissolution, approximately 150 ng of Sr was loaded onto outgassed single Re filament assemblies, along with 1 µl of a $TaCl_5$ solution. ^{88}Sr/^{84}Sr, ^{87}Sr/^{84}Sr and ^{86}Sr/^{84}Sr ratios were collected in static mode. At least 140 cycles were used for the data reduction, which includes input of corresponding ^{87}Sr/^{86}Sr ratios. All ^{88}Sr/^{86}Sr ratios are reported in the same δ notation relative to NBS 987. At least six NBS 987 standards and two IAPSO seawater standards were analysed every 20 samples. During the period of study, repeated analyses yielded $\delta^{88/86}Sr_{NBS987}$ = 0.000 ± 0.005‰ ($2\sigma_{SEM}$, n = 18) and $\delta^{88/86}Sr_{IAPSO}$ = 0.398 ± 0.008‰ (±$2\sigma_{SEM}$, n = 6), consistent with long-term laboratory results and previously published data[24,40]. Therefore the uncertainty assigned to the double-spike TIMS $\delta^{88/86}$Sr analyses is ±0.020‰ ($2\sigma_{SD}$). Notably, replicate measurements by MC-ICP-MS and TIMS are consistent (Supplementary Table 1).

Petrographic and geochemical screening for diagenetic alteration

Petrography. As well as carefully avoiding any visible fractures, veins and secondary cements, all carbonate samples were petrographically screened under optical and cathodoluminescence microscopy before selection for geochemical analysis. Under optical microscopy, any samples showing signs of recrystallization, late-stage cementation, fracturing and veining were avoided; under cathodoluminescence microscopy, any samples that showed layered or spotty non-uniform luminescence were also avoided[41]. For the 437 Precambrian samples that were examined, 276 samples were petrographically identified as well preserved, and then micro-drilled, sequentially leached and analysed for their elemental concentrations.

Trace elements. Strict geochemical screening criteria were applied to select samples that had undergone the least diagenetic alteration and had the highest likelihood of recording seawater Sr isotope compositions. Sr and Mn concentrations were first used as diagenetic indicators (for example, refs. [42–44]). The Sr partition coefficient is kinetically controlled and diagenetic alteration (that is, recrystallization, dissolution and other rock–fluid interactions) is often associated with slow reaction rates, leading to Sr loss[43,45,46] and limited reuptake owing to distribution coefficients an order of magnitude lower for calcite (0.14)[47,48] than for aragonite (1.12)[49]. Subsurface fluids have higher Mn concentrations and diagenetically altered carbonates tend to have high Mn (ref. [50]). In this study, we excluded any samples with low Sr and high Mn concentrations (that is, Sr < 100 ppm and Mn > 300 ppm) and only selected samples with high Sr and low Mn according to these criteria. The selected samples yielded low Mn/Sr ratios, with mean values of 0.4 for calcites and 1.2 for dolomites. Samples with high Pb concentrations (that is, Pb > 1 ppm) were excluded to avoid potential diagenetic overprinting by metal-rich, anoxic fluids[51]. Given high concentrations of Rb and Ti in clays and other detrital minerals, we also used Rb and Ti concentrations as indicators of detrital contamination[52–54]. Furthermore, ^{87}Sr can be generated by β decay of ^{87}Rb with a half-life of 48.8 × 10^9 years and Rb excesses from detrital materials in Precambrian rocks could affect radiogenic Sr isotope ratios. Thus, we further avoided any samples with high Ti and Rb concentrations (that is, Ti > 5 ppm and Rb > 1 ppm). After application of these rigorous geochemical and petrographic screening criteria, 115 out of the 276 samples that had met petrographic screening criteria were ultimately selected for Sr isotope analyses.

Radiogenic Sr (^{87}Sr/^{86}Sr). We further used ^{87}Sr/^{86}Sr as another diagenetic indicator. Although early diagenesis (that is, neomorphism and recrystallization) only minimally alters bulk carbonate ^{87}Sr/^{86}Sr values[45,46,55], exogenous Sr input during late-stage diagenesis could potentially exert a much greater influence on bulk carbonate ^{87}Sr/^{86}Sr compositions[22,54,56]. Late-stage fluids—such as brine, groundwater and meteoric fluids—that interact with Rb-rich clay minerals, feldspars and mica in siliciclastic sediments and crystalline basement rocks often impart high ^{87}Sr/^{86}Sr ratios to sediments and rocks that they subsequently contact (for example, refs. [57–59]), and extensive interactions with these fluids could strongly shift carbonate ^{87}Sr/^{86}Sr ratios (for example, refs. [22,56]). To ensure that our selected samples provide a reliable seawater Sr archive, we compared the measured ^{87}Sr/^{86}Sr ratio of each sample to contemporaneous seawater ^{87}Sr/^{86}Sr records—the lowest 10% of the ^{87}Sr/^{86}Sr ratios at each age in the Precambrian Marine Carbonate Isotope Database (PMCID)[23]) were bootstrap-resampled (n = 1,000) and any measured samples with ^{87}Sr/^{86}Sr ratios outside the resampling range were rejected. As shown in Extended Data Fig. 1, the ^{87}Sr/^{86}Sr ratios of selected Precambrian samples closely follow the local regressed value at each age (locally estimated scatterplot smoothing (LOESS) fit of the lowest 10% PMCID values), validating this selection criterion. This criterion filtered out most of the dolomite samples, as well as a subset of the calcite samples, given their elevated ^{87}Sr/^{86}Sr ratios (diamonds and crosses in Extended Data Figs. 1 and 2). As shown in Extended Data Fig. 3, the $\delta^{88/86}$Sr values of studied dolomites are, on the whole, negatively correlated with ^{87}Sr/^{86}Sr values (that is, lower $\delta^{88/86}$Sr values are associated with higher ^{87}Sr/^{86}Sr values), indicating systematic overprinting of Sr isotope compositions for these dolomites. One possible explanation is that exogenous fluids affecting these rocks became enriched in radiogenic Sr (higher ^{87}Sr/^{86}Sr values; for example, refs. [22,56]) and isotopically depleted stable Sr (lower $\delta^{88/86}$Sr values) during interactions with younger sedimentary successions and associated fluids. Therefore the dolomite samples screened here with elevated ^{87}Sr/^{86}Sr ratios are classified as diagenetically altered using the criteria used by this study. The lower $\delta^{88/86}$Sr values of these excluded samples fall closer to the bulk silicate Earth value, compared with the best-preserved calcite samples (Extended Data Fig. 1 and dolomites in Extended Data Fig. 2). Various studies have suggested that some Precambrian dolomites may have a primary (for example, synsedimentary or early diagenetic) origin (for example, ref. [60]). Notably, we observed that some of the dolomites analysed in this study—particularly those with well-preserved petrographic fabrics and low ^{87}Sr/^{86}Sr ratios—exhibit elevated $\delta^{88/86}$Sr values that are comparable with those of contemporaneous calcites (Extended Data Figs. 1–3). Because dolomite precipitation is expected to be characterized by similar (and kinetically controlled) $\delta^{88/86}$Sr fractionation as that observed for other carbonate minerals[25], we infer that well-preserved primary or syndepositional Precambrian dolomites may, like well-preserved calcites, provide robust archives of elevated seawater $\delta^{88/86}$Sr values (reflecting, in turn, high seawater Ω_{carb}). In short, we regard carbonates that passed both petrographic and geochemical (Sr, Mn, Pb, Ti, Rb) screening criteria, and with ^{87}Sr/^{86}Sr ratios closely aligned with contemporaneous seawater ^{87}Sr/^{86}Sr values (calcite, n = 59), to represent the best-preserved samples—and among those samples analysed for this study, most were well-preserved non-skeletal calcites. These $\delta^{88/86}$Sr values are plotted with published $\delta^{88/86}$Sr data from other studies[14,15,24,25,61,62] in Fig. 1 and Extended Data Fig. 1.

Supporting elemental and isotopic results. We evaluated the statistical correlations between the $\delta^{88/86}$Sr values of the best-preserved calcites and different elemental parameters. We used a standardized major axis (SMA) linear regression model to evaluate the statistical significance of each correlation. Unlike the ordinary least squares model, the SMA model considers that neither x nor y variables control correlations and that both have error. As shown in Extended Data Fig. 4, no statistically significant correlations exist between Sr isotope ratios and elemental concentrations, which supports the fidelity of the generated ^{87}Sr/^{86}Sr and $\delta^{88/86}$Sr data. In light of these data, we interpret the screened $\delta^{88/86}$Sr values to faithfully reflect the $\delta^{88/86}$Sr signatures of primary carbonates that precipitated from contemporaneous seawater.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data are available in the main text or the Supplementary Information. All data are also reposited in EarthChem (https://doi.org/10.26022/IEDA/112713).

Code availability

We used the open-source language R (version 4.1.1) to analyse the measured data, analyse the EarthChem (http://portal.earthchem.org/) and Macrostrat (https://macrostrat.org/#api) datasets, and generate all plots. All equations for the mass-balance model are listed in the Supplementary Information and all associated code is deposited on GitHub (https://github.com/julianwangnwu/carbonatefactoryevolution).

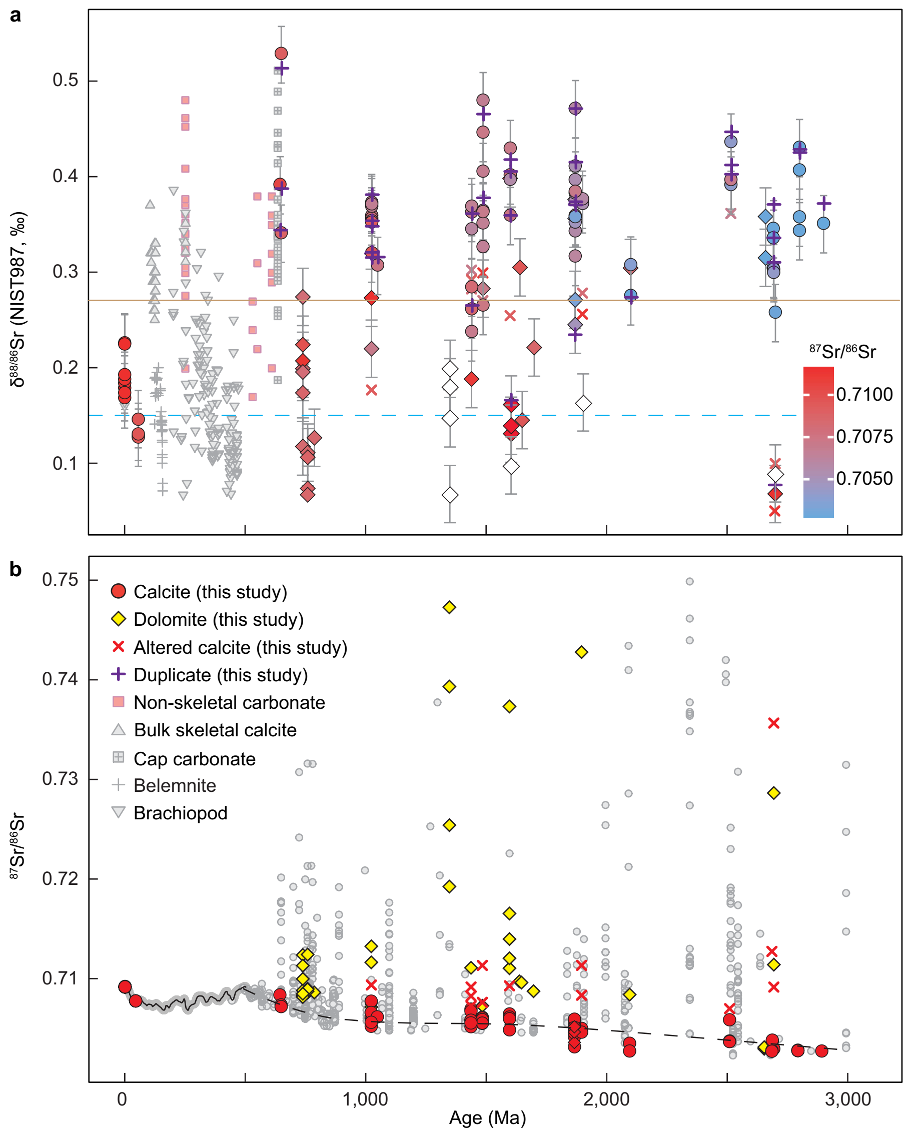
35. Krissansen-Totton, J., Buick, R. & Catling, D. C. A statistical analysis of the carbon isotope record from the Archean to Phanerozoic and implications for the rise of oxygen. *Am. J. Sci.* **315**, 275–316 (2015).
36. Liu, C., Wang, Z. & Raub, T. D. Geochemical constraints on the origin of Marinoan cap dolostones from Nuccaleena Formation, South Australia. *Chem. Geol.* **351**, 95–104 (2013).
37. Wang, J., Asael, D., Planavsky, N. J. & Tarhan, L. G. An investigation of factors affecting high-precision Sr isotope analyses (^{87}Sr/^{86}Sr and $\delta^{88/86}$Sr) by MC-ICP-MS. Preprint at https://arxiv.org/abs/2111.02942 (2021).
38. Ohno, T. & Hirata, T. Simultaneous determination of mass-dependent isotopic fractionation and radiogenic isotope variation of strontium in geochemical samples by multiple collector-ICP-mass spectrometry. *Anal. Sci.* **23**, 1275–1280 (2007).
39. Ma, J. L. et al. Precise measurement of stable ($\delta^{88/86}$Sr) and radiogenic (^{87}Sr/^{86}Sr) strontium isotope ratios in geological standard reference materials using MC-ICP-MS. *Chin. Sci. Bull.* **58**, 3111–3118 (2013).
40. Andrews, M. G., Jacobson, A. D., Lehn, G. O., Horton, T. W. & Craw, D. Radiogenic and stable Sr isotope ratios (^{87}Sr/^{86}Sr, $\delta^{88/86}$Sr) as tracers of riverine cation sources and biogeochemical cycling in the Milford Sound region of Fiordland, New Zealand. *Geochim. Cosmochim. Acta* **173**, 284–303 (2016).
41. Machel, H. G. in *Cathodoluminescence in Geosciences* 271–301 (Springer, 2000).
42. Bathurst, R. G. C. *Carbonate Sediments and Their Diagenesis* (Elsevier, 1972).
43. Brand, U. & Veizer, J. Chemical diagenesis of a multicomponent carbonate system; 1, trace elements. *J. Sediment. Res.* **50**, 1219–1236 (1980).
44. Anderson, T. F. & Arthur, M. A. in *Stable Isotopes in Sedimentary Geology* (eds Arthur, M. A., Anderson, T. F., Kaplan, I. R., Veizer, J. & Land, L. S.) (SEPM Society for Sedimentary Geology, 1983).
45. Richter, F. M. & DePaolo, D. J. Diagenesis and Sr isotopic evolution of seawater using data from DSDP 590B and 575. *Earth Planet. Sci. Lett.* **90**, 382–394 (1988).
46. Richter, F. M. & Liang, Y. The rate and consequences of Sr diagenesis in deep-sea carbonates. *Earth Planet. Sci. Lett.* **117**, 553–565 (1993).
47. Holland, H., Holland, H. & Munoz, J. The coprecipitation of cations with $CaCO_3$—II. The coprecipitation of Sr^{+2} with calcite between 90° and 100°C. *Geochim. Cosmochim. Acta* **28**, 1287–1301 (1964).
48. Katz, A., Sass, E., Starinsky, A. & Holland, H. D. Strontium behavior in the aragonite-calcite transformation: an experimental study at 40–98°C. *Geochim. Cosmochim. Acta* **36**, 481–496 (1972).
49. Kinsman, D. J. J. & Holland, H. D. The co-precipitation of cations with $CaCO_3$—IV. The co-precipitation of Sr^{2+} with aragonite between 16° and 96°C. *Geochim. Cosmochim. Acta* **33**, 1–17 (1969).
50. Derry, L. A., Kaufman, A. J. & Jacobsen, S. B. Sedimentary cycling and environmental change in the Late Proterozoic: evidence from stable and radiogenic isotopes. *Geochim. Cosmochim. Acta* **56**, 1317–1329 (1992).
51. Cruse, A. M. & Lyons, T. W. Trace metal records of regional paleoenvironmental variability in Pennsylvanian (Upper Carboniferous) black shales. *Chem. Geol.* **206**, 319–345 (2004).
52. Sageman, B. B. et al. A tale of shales: the relative roles of production, decomposition, and dilution in the accumulation of organic-rich strata, Middle-Upper Devonian, Appalachian basin. *Chem. Geol.* **195**, 229–273 (2003).
53. Piper, D. Z. & Calvert, S. E. A marine biogeochemical perspective on black shale deposition. *Earth-Sci. Rev.* **95**, 63–96 (2009).
54. Banner, J. L. & Hanson, G. N. Calculation of simultaneous isotopic and trace element variations during water-rock interaction with applications to carbonate diagenesis. *Geochim. Cosmochim. Acta* **54**, 3123–3137 (1990).
55. Baker, P. A., Gieskes, J. M. & Elderfield, H. Diagenesis of carbonates in deep-sea sediments; evidence from Sr/Ca ratios and interstitial dissolved Sr^{2+} data. *J. Sediment. Res.* **52**, 71–82 (1982).
56. Chaudhuri, S. & Clauer, N. Strontium isotopic compositions and potassium and rubidium contents of formation waters in sedimentary basins: clues to the origin of the solutes. *Geochim. Cosmochim. Acta* **57**, 429–437 (1993).
57. Stueber, A. M., Pushkar, P. & Hetherington, E. A. A strontium isotopic study of Smackover brines and associated solids, southern Arkansas. *Geochim. Cosmochim. Acta* **48**, 1637–1649 (1984).
58. McNutt, R. H., Frape, S. K., Fritz, P., Jones, M. G. & MacDonald, I. M. The ^{87}Sr/^{86}Sr values of Canadian Shield brines and fracture minerals with applications to groundwater mixing, fracture history, and geochronology. *Geochim. Cosmochim. Acta* **54**, 205–215 (1990).
59. McNutt, R. H., Frape, S. K. & Fritz, P. Strontium isotopic composition of some brines from the Precambrian Shield of Canada. *Chem. Geol.* **46**, 205–215 (1984).
60. Wilcots, J., Gilbert, P. U. P. A. & Bergmann, K. D. Nanoscale crystal fabric of primary Ediacaran dolomite. Preprint at *ESS Open Archive* https://doi.org/10.1002/essoar.10507750.1 (2021).
61. Ohno, T., Komiya, T., Ueno, Y., Hirata, T. & Maruyama, S. Determination of ^{88}Sr/^{86}Sr mass-dependent isotopic fractionation and radiogenic isotope variation of ^{87}Sr/^{86}Sr in the Neoproterozoic Doushantuo Formation. *Gondwana Res.* **14**, 126–133 (2008).
62. Sawaki, Y. et al. Sr isotope excursion across the Precambrian–Cambrian boundary in the Three Gorges area, South China. *Gondwana Res.* **14**, 134–147 (2008).
63. Pearce, C. R. et al. Reassessing the stable ($\delta^{88/86}$Sr) and radiogenic (^{87}Sr/^{86}Sr) strontium isotopic composition of marine inputs. *Geochim. Cosmochim. Acta* **157**, 125–146 (2015).
64. McArthur, J. M., Howarth, R. J. & Shields, G. A. in *The Geologic Time Scale* (eds Gradstein, F. M., Ogg, J. G., Schmitz, M. & Ogg, G. M.) 127–144 (Elsevier, 2012).

Acknowledgements We thank S. Nicolescu, B. Kalderon-Asael and Y. Wang for facilitating access to the Yale Peabody Museum and Woods Hole Oceanographic Institution collections and for assistance with sample selection; D. Asael for assistance with MC-ICP-MS method development; R. P. Reid and E. P. Suosaari for access to the Hamelin Pool stromatolite samples; S. Ye for assistance with the Macrostrat database; and D. Schrag, M. Arthur, K. Bergmann, Z. Zhang and Y. Cui for helpful discussions. This study is supported by an Agouron Geobiology Postdoctoral Fellowship to J.W. and National Aeronautics and Space Administration Interdisciplinary Consortia for Astrobiology Research grant (NNA15BB03A) to N.J.P.

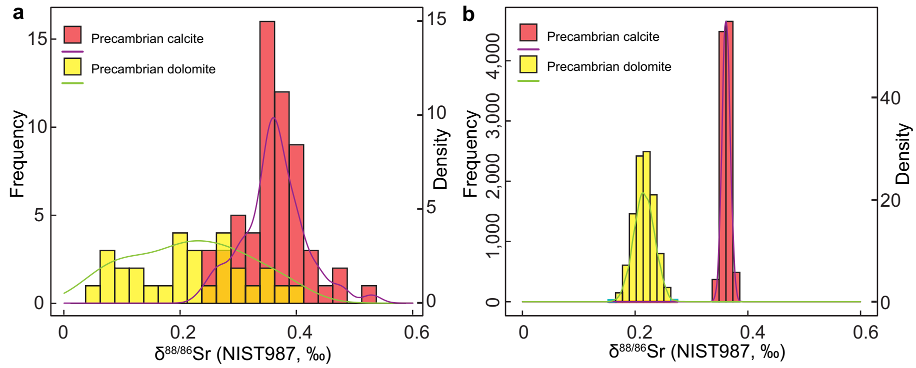
Author contributions J.W., L.G.T. and N.J.P. conceived the study and acquired funding. J.W., L.G.T., A.D.J. and N.J.P. developed the methodology. J.W. performed mass spectrometry analyses. J.W. and L.G.T. conducted the statistical analyses. J.W. and L.G.T. wrote the paper, with input from A.D.J., A.M.O. and N.J.P. J.W., L.G.T., A.D.J., A.M.O. and N.J.P. all contributed to the interpretation of the results and editing the manuscript.

Competing interests The authors declare no competing interests.

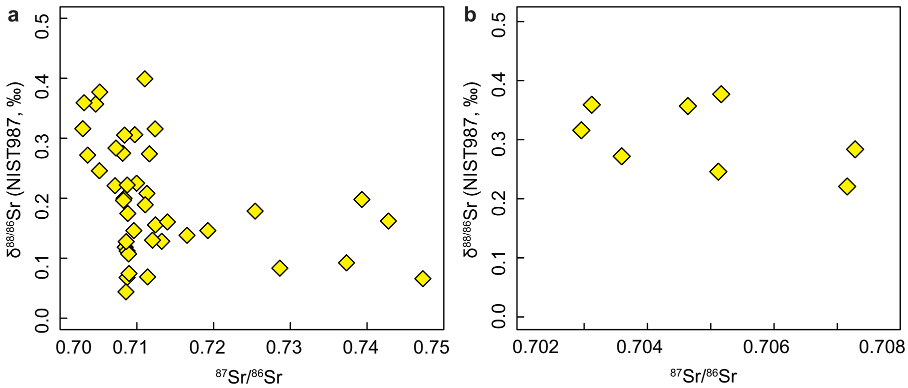
Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-05654-5.
Correspondence and requests for materials should be addressed to Jiuyuan Wang or Lidya G. Tarhan.
Peer review information *Nature* thanks Adina Paytan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



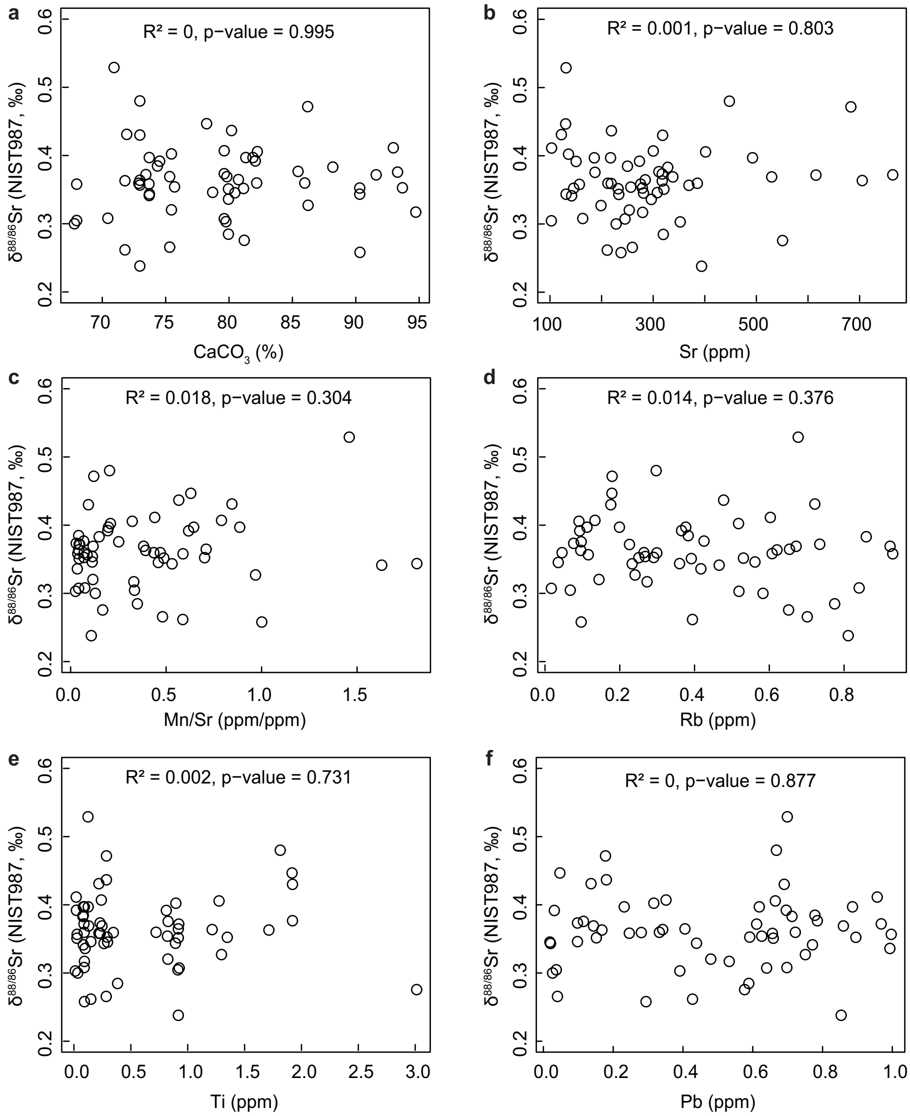
Extended Data Fig. 1 | Marine carbonate strontium isotope records through the history of the Earth. a, Summary of $\delta^{88/86}$Sr values measured in marine calcites and dolomites. New data from this study (n = 139) are colour-contoured to indicate corresponding radiogenic Sr isotope ratios (^{87}Sr/^{86}Sr) generated from the same samples: circles, calcite; diamonds, dolomite; ×, calcite with abnormally high ^{87}Sr/^{86}Sr ratios. Error bars represent the long-term external reproducibility of $\delta^{88/86}$Sr ($2\sigma_{SD}$ = ±0.03‰, n = 273). Purple crosses denote duplicate measurements of the same sample (see Methods for description of duplicate strategy). The gold line illustrates the $\delta^{88/86}$Sr value of bulk silicate Earth (0.27‰)[27]. The dashed blue line represents the $\delta^{88/86}$Sr value of modern marine carbonate[63]. Other symbols represent published data from other studies (n = 299; see Methods): pink squares, non-skeletal carbonate; grey gridded squares, cap carbonate; grey triangles, bulk skeletal calcite; grey crosses, belemnite; grey inverted triangles, brachiopod. **b**, Marine carbonate ^{87}Sr/^{86}Sr ratios. New data from this study are denoted by coloured symbols: circles, calcite; diamonds, dolomite; ×, calcite with abnormally high ^{87}Sr/^{86}Sr ratios. The grey circles represent Precambrian carbonate ^{87}Sr/^{86}Sr records (n = 1,494)[23]. The dashed curve denotes the LOESS fit of the lowest 10% of Precambrian ^{87}Sr/^{86}Sr ratios[23] and the solid curve denotes the LOESS fit of the Phanerozoic ^{87}Sr/^{86}Sr record[64].



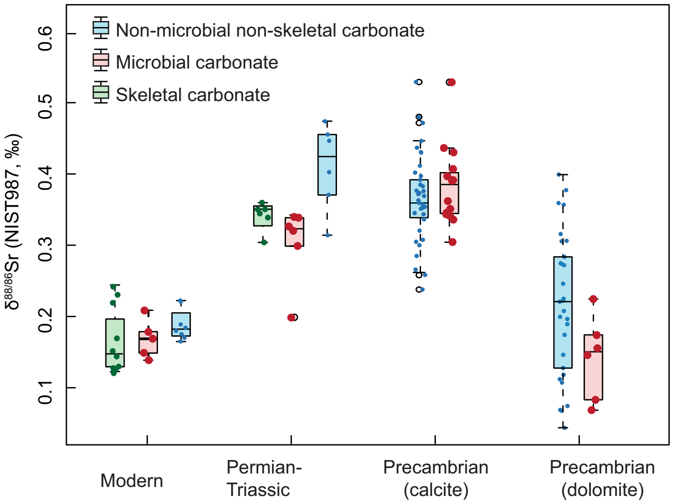
Extended Data Fig. 2 | Histograms of measured and bootstrap-resampled Precambrian calcite and dolomite $\delta^{88/86}$Sr values. a, Measured Precambrian calcite (red) and dolomite (yellow) $\delta^{88/86}$Sr values. **b**, Bootstrap-resampled (n = 10,000) Precambrian calcite (red) and dolomite (yellow) $\delta^{88/86}$Sr values. All Precambrian calcite and dolomite $\delta^{88/86}$Sr values are from this study. The purple and green curves represent density distributions of $\delta^{88/86}$Sr in Precambrian calcites (purple, n = 72) and dolomites (green, n = 43).



Extended Data Fig. 3 | The relationship between $\delta^{88/86}$Sr and ^{87}Sr/^{86}Sr in analysed dolomites. **a**, The stable and radiogenic Sr isotope relationship for all analysed dolomites (n = 43). A SMA regression model yields R^2 = 0.223 and P = 0.001. **b**, The stable and radiogenic Sr isotopic values of less-altered dolomite samples from this dataset, that is, samples characterized by ^{87}Sr/^{86}Sr values less than 0.708, the inferred value of Ediacaran seawater[64].



Extended Data Fig. 4 | Cross-plots of $\delta^{88/86}$Sr versus different elemental contents and ratios. a, $\delta^{88/86}$Sr versus $CaCO_3$ weight percentage (wt%). Carbonate wt% calculated using calcium content assuming stoichiometric $CaCO_3$. **b**, $\delta^{88/86}$Sr versus Sr contents. **c**, $\delta^{88/86}$Sr versus Mn/Sr. **d**, $\delta^{88/86}$Sr versus Rb contents. **e**, $\delta^{88/86}$Sr versus Ti contents. **f**, $\delta^{88/86}$Sr versus Pb contents. $\delta^{88/86}$Sr values are given in ‰, normalized to NIST 987; all elemental concentrations are in ppm except when noted otherwise. An SMA regression model was used to evaluate the statistical significance of each correlation. R^2 and P-values are listed at the top of each panel. No statistical correlations are observed at the significance level of 0.01.



Extended Data Fig. 5 | Box plot of $\delta^{88/86}$Sr values for skeletal (green), microbial (red) and non-skeletal, non-microbial carbonate (blue) for modern, Permian–Triassic and Precambrian deposits. In the box plot, the centre line represents the median of the data (50th percentile), box limits represent the upper and lower quartiles (75th and 25th percentiles), whiskers represent 1.5 times the interquartile range, blank points represent outliers and coloured points represent all data. These data indicate that the notably higher $\delta^{88/86}$Sr values characterizing Precambrian calcites cannot be attributed to differences between microbial and non-microbial pathways of carbonate precipitation and that, similarly, the shift between elevated Precambrian and less-elevated Phanerozoic $\delta^{88/86}$Sr values cannot be readily attributed to differences between skeletal and non-skeletal pathways of carbonate precipitation. The modern $\delta^{88/86}$Sr records are from Stevenson et al.[20] (skeletal $n = 10$) and this study (microbial $n = 5$; non-skeletal, non-microbial $n = 8$); the Permian–Triassic $\delta^{88/86}$Sr records (skeletal $n = 6$; microbial $n = 8$; non-skeletal, non-microbial $n = 20$) are from Wang et al.[24]; the Precambrian calcite (microbial calcite $n = 12$; non-skeletal, non-microbial calcite $n = 47$) and dolomite (microbial dolomite $n = 6$; non-skeletal, non-microbial dolomite $n = 37$) $\delta^{88/86}$Sr records are from this study.

nature portfolio

Corresponding author(s): Lidya Tarhan and Jiuyuan Wang

Last updated by author(s): Jul 23, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used to collect these data.
Data analysis	We used the open-source language R (version 4.1.1) to analyse the measured data, EarthChem (http://portal.earthchem.org/) and Macrostrat (https://macrostrat.org/#api) datasets, and generate all plots. All equations for the mass balance model are listed in the Supplementary Information.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data are available in the main text or the Supplementary Information. All data are additionally reposited in EarthChem (data to be reposited and DOI to be populated upon acceptance of the manuscript). All analysed samples are hosted in public repositories (i.e., Yale Peabody Museum Mineralogy and Meteoritics

Collection, Woods Hole Oceanographic Institution Seafloor Sample Repository, and International Ocean Discovery Program Gulf Coast Repository) and their unique sample identifiers are provided in the Supplementary Information.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	No human research participants were involved in this study.
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This is an analytical and modeling-based geochemical study. Geological samples were analysed to determine their strontium isotope composition and the corresponding data were placed into the context of a global strontium isotope mass balance model to make inferences about the long-term evolution of the global carbon cycle and the marine carbonate factory.
Research sample	We used geologic samples of petrographically screened carbonates from public repositories for all strontium isotope analyses.
Sampling strategy	Samples were selected based on their petrographic and geochemical characteristics—targeting samples with limited alteration. The sample number was determined by sample availability and the generation of statistically significant trends through time in the data.
Data collection	Strontium isotope data were generated and collected via MC-ICP-MS and TIMS, following standard previously published methods, by Jiuyuan Wang, with supervision from Andrew Jacobson, Noah Planavsky and Lidya Tarhan. Replicates analyses were performed and are reported in the Supplementary Information.
Timing and spatial scale	Data were collected as continuously as possible, given mass spectrometer availability, from 2020 to early 2022.
Data exclusions	No data were excluded from the analyses unless explicitly stated in the text. All analysed data are reported.
Reproducibility	Reproducibility is based on duplicate analysis of geostandards and samples. Error is reported in the text.
Randomization	Sample groupings were based on standard divisions of geologic time widely employed by the international geological community.
Blinding	All samples were given a random analysis number for sample preparation and sample analysis.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Tropical deforestation causes large reductions in observed precipitation

https://doi.org/10.1038/s41586-022-05690-1

Received: 14 April 2022

Accepted: 15 December 2022

Published online: 1 March 2023

Open access

Check for updates

C. Smith[1]✉, J. C. A. Baker[1] & D. V. Spracklen[1]

Tropical forests play a critical role in the hydrological cycle and can influence local and regional precipitation[1]. Previous work has assessed the impacts of tropical deforestation on precipitation, but these efforts have been largely limited to case studies[2]. A wider analysis of interactions between deforestation and precipitation—and especially how any such interactions might vary across spatial scales—is lacking. Here we show reduced precipitation over deforested regions across the tropics. Our results arise from a pan-tropical assessment of the impacts of 2003–2017 forest loss on precipitation using satellite, station-based and reanalysis datasets. The effect of deforestation on precipitation increased at larger scales, with satellite datasets showing that forest loss caused robust reductions in precipitation at scales greater than 50 km. The greatest declines in precipitation occurred at 200 km, the largest scale we explored, for which 1 percentage point of forest loss reduced precipitation by 0.25 ± 0.1 mm per month. Reanalysis and station-based products disagree on the direction of precipitation responses to forest loss, which we attribute to sparse in situ tropical measurements. We estimate that future deforestation in the Congo will reduce local precipitation by 8–10% in 2100. Our findings provide a compelling argument for tropical forest conservation to support regional climate resilience.

Tropical forests play an important role in moderating local, regional and global climate through their impact on energy, water and carbon cycles[3]. Crucially, tropical forests control local-to-regional rainfall patterns[1,2]. Evapotranspiration from tropical forests is a strong driver of regional precipitation[4,5] contributing up to 41% of basin mean rainfall over the Amazon and up to 50% over the Congo[6]. Evergreen tropical forests are dependent on high annual rainfall for their survival and productivity[7], and forest–rainfall feedbacks have been highlighted as an important determinant of tropical forest stability[4,5,8], amid concerns that the exacerbating impacts of droughts and deforestation could threaten their viability[9].

Rapid loss of forests is occurring across the tropics[10]. Tropical deforestation warms the climate at local-to-global scales by changing the surface energy balance and through emissions of carbon dioxide[3]. The impact of tropical deforestation on precipitation is less certain with a range of processes operating at different scales. Small-scale deforestation over the southern Amazon has been shown to increase precipitation frequency[11,12] owing to thermally[13] and dynamically[12] induced circulations. At larger scales, deforestation reduces precipitation recycling leading to a reduction in precipitation[1,14]. Over Indonesia, deforestation has been linked to declining precipitation[15], and exacerbation of El Niño impacts[16]. Global and regional climate models predict annual precipitation declines of 8.1 ± 1.4% for large-scale Amazonian deforestation by 2050 (ref. [17]), but an observational study of the impacts of tropical deforestation on precipitation across spatial scales is lacking.

Here we present a pan-tropical assessment of the impact of forest loss on precipitation based on measurements. We use a satellite dataset of forest cover change over the period 2003–2017 to identify areas of forest loss, with a focus on evergreen broadleaf forests of the Amazon, Congo and Southeast Asia (SEA; Fig. 1). To provide a robust assessment of the impacts of deforestation on precipitation, we analysed 18 different precipitation datasets, including satellite ($n = 10$), station-based ($n = 4$) and reanalysis ($n = 4$) products (Extended Data Table 1). We compared the precipitation change over pixels experiencing forest loss with neighbouring pixels that had experienced less forest loss (Methods). This comparison against neighbouring pixels that will have experienced similar climate change focuses our analysis on the changes due to forest loss. To explore the impact of forest loss across scales, we analysed the impacts of forest loss on coincident precipitation at a series of spatial resolutions ranging from roughly 5 km to 200 km (0.05°, 0.1°, 0.25°, 0.5°, 1.0° and 2.0°).

Precipitation response to forest loss

Observed precipitation responses to tropical forest loss across multiple spatial scales and precipitation products are presented in Fig. 2. Satellite-based precipitation datasets suggest that tropical forest loss causes statistically significant ($P < 0.05$) declines in median annual mean precipitation at all scales analysed. At larger scales (>0.5°), reductions exceed 0.03 mm per month for each percentage point loss of forest cover (Fig. 2d–f). The largest changes are observed at the 2.0° scale (approximately 220 km at the Equator; Fig. 2f), for which each percentage point reduction in forest cover causes 0.25 ± 0.1 mm per month reduction in annual precipitation.

Analysis of precipitation change as a function of forest loss confirms larger reductions in precipitation for larger reductions in forest cover

[1]School of Earth and Environment, University of Leeds, Leeds, UK. ✉e-mail: ee13c2s@leeds.ac.uk

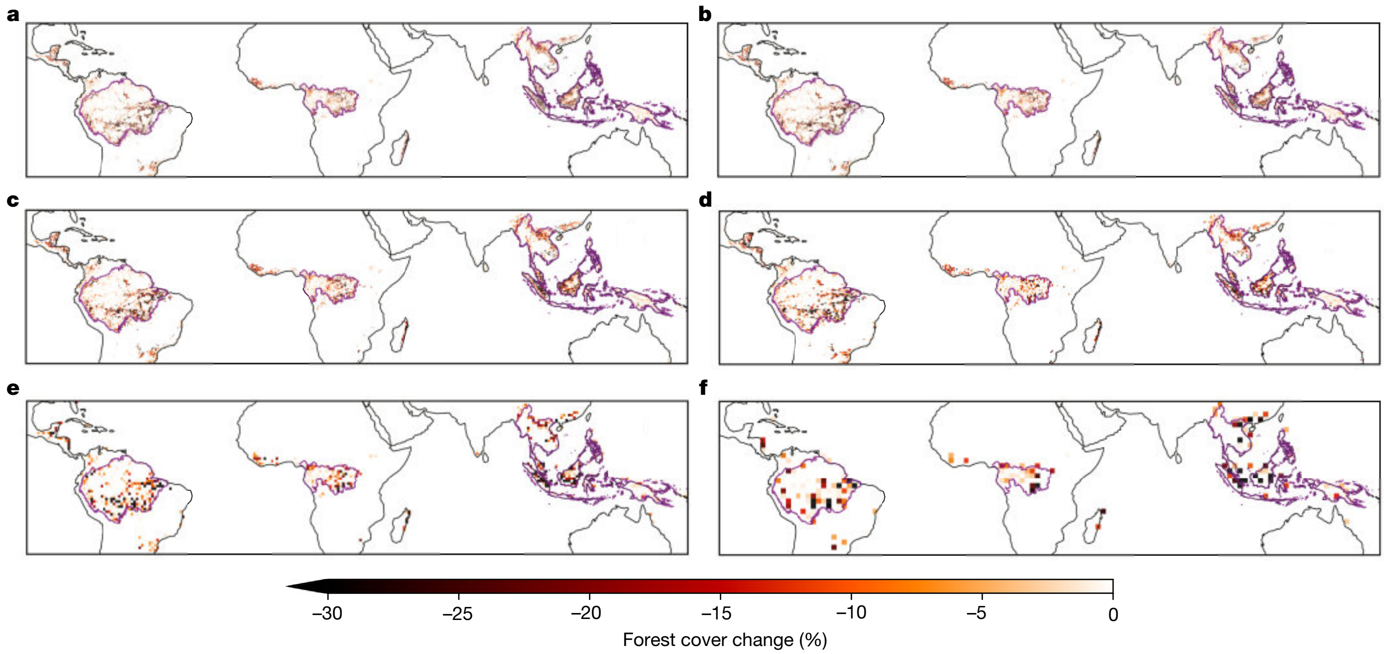


Fig. 1 | Tropical evergreen broadleaf forest cover loss from 2003 to 2017. **a**–**f**, Forest cover change at 0.05° (**a**), 0.1° (**b**), 0.25° (**c**), 0.5° (**d**), 1.0° (**e**) and 2.0° (**f**) resolution. The Amazon Basin, Congo Basin and SEA regions used in this study are outlined in purple. Maps of the different regions generated using Cartopy and Natural Earth[51]. Forest loss data from ref. [10].

(Extended Data Fig. 1), although with considerable variability, as seen in the modelled response[18]. Observed reductions in precipitation are consistent across satellite datasets, with all ten satellite precipitation products agreeing on the sign of the rainfall response at 2° over the tropics (Extended Data Fig. 2). At the 2° scale, significant ($P < 0.05$) reductions in annual mean precipitation with forest loss were observed across all tropical regions (Fig. 2). Reductions in precipitation at 2° based on satellite datasets ranged from 0.48 ± 0.36 mm per month in

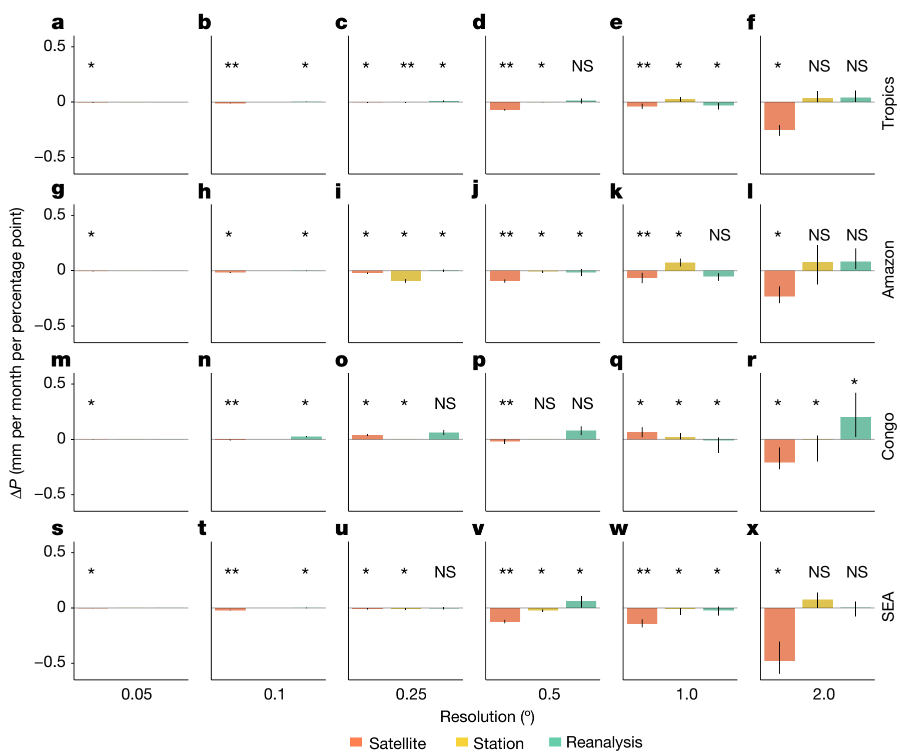


Fig. 2 | Reductions in precipitation over regions of tropical forest loss. **a**–**r**, Bars indicate the median absolute change in annual precipitation (millimetres per month) per percentage point of forest loss over 2003 to 2017 in each region (tropics (**a**–**f**), Amazon (**g**–**l**), Congo (**m**–**r**), SEA (**s**–**x**)) for each precipitation dataset category (satellite, station and reanalysis). Results are shown for forest loss scales of 0.05° (**a**,**g**,**m**,**s**), 0.1° (**b**,**h**,**n**,**t**), 0.25° (**c**,**i**,**o**,**u**), 0.5° (**d**,**j**,**p**,**v**), 1.0° (**e**,**k**,**q**,**w**) and 2.0° (**f**,**l**,**r**,**x**). Statistically significant (*$P < 0.05$; **$P < 0.01$) and nonsignificant (NS) differences in changes in mean precipitation (calculated as a multi-annual mean over 2003–2007 compared with 2013–2017) over deforested regions compared with control regions are indicated. Error bars show ±1 standard error from the mean. Datasets used in this analysis are detailed in Extended Data Table 1. Δ*P*, precipitation change.

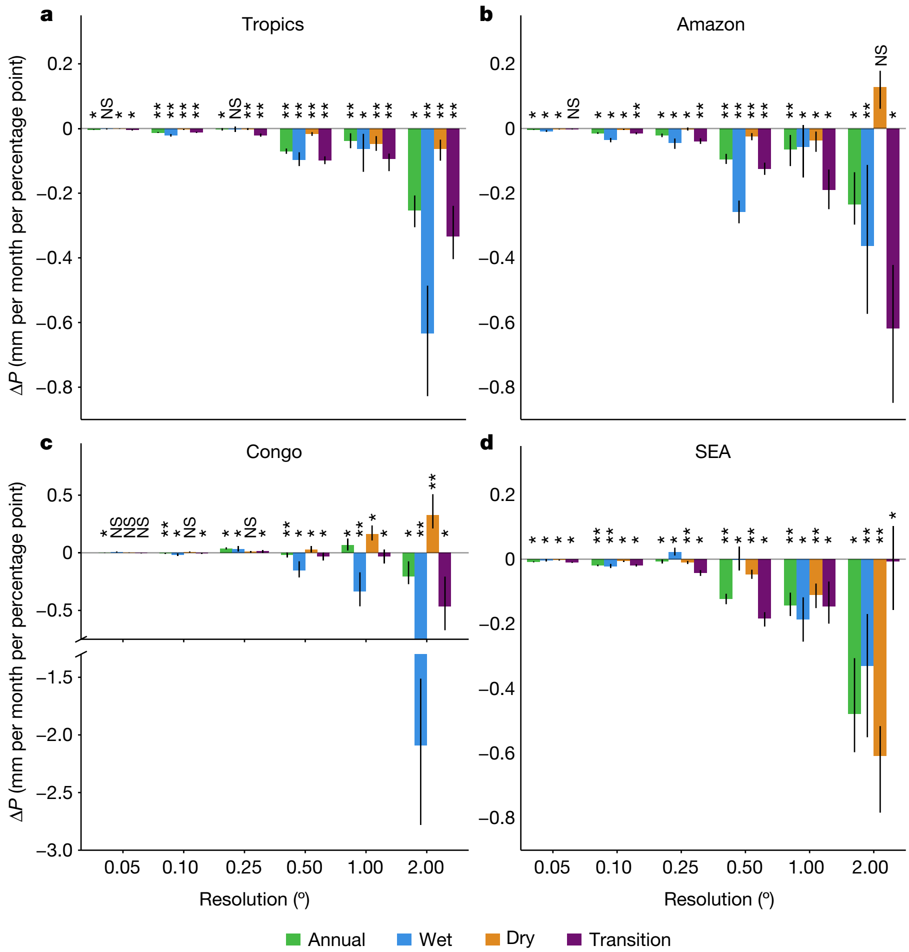


Fig. 3 | Changes in seasonal precipitation due to forest loss. a–d, Bars indicate the median change in precipitation (millimetres per month) per percentage point forest cover loss for satellite datasets during 2003–2017 for tropics (**a**), Amazon (**b**), Congo (**c**) and SEA (**d**). Error bars indicate ±1 standard error from the mean. Statistically significant (*$P < 0.05$; **$P < 0.01$) and nonsignificant (NS) differences in changes in mean precipitation over deforested regions compared with controls are indicated. Results are shown for the wettest 3 months (wet), the driest 3 months (dry) and the transition months (remaining 6 months). Datasets used in this analysis detailed in Extended Data Table 1.

SEA to 0.23 ± 0.12 mm per month in the Amazon, and 0.21 ± 0.19 mm per month in the Congo for each percentage point loss in forest cover, with at least 8 out of 10 satellite datasets agreeing on the sign of the response within each region (Extended Data Fig. 2). In SEA, it has been suggested that proximity to the ocean and the replacement of tropical forest with plantations as opposed to pasture or cropland may cause reduced sensitivity of precipitation to deforestation[1]. Our analysis suggests that forest loss in SEA causes reductions in precipitation consistent with or greater than reductions in precipitation in the Amazon and Congo.

Station-based datasets and reanalysis products exhibit contrasting annual mean precipitation responses to deforestation at 2.0° (Fig. 2). Across the tropics, station-based and reanalysis datasets showed no statistically significant changes in annual mean precipitation due to forest loss (Fig. 2f), and there was little agreement with satellite datasets at the regional scale (Fig. 2l,r,x), with some non-satellite precipitation products showing small increases in annual mean precipitation due to forest loss. Sparse in situ measurements across the tropics, particularly over regions of forest loss, mean that station-based datasets provide a weak constraint on precipitation changes. A comparison of station-based precipitation datasets revealed higher levels of uncertainty in the tropics, including the Amazon[19]. In regions of sparse data such as tropical forests[20], interpolation methods may mask precipitation changes driven by forest loss. Reanalysis products, which are numerical models constrained by empirical data, are also expected to be less reliable in regions where in situ data are limited[21]. Our results indicate that precipitation data based on satellite remote-sensing measurements may have an advantage over tropical forest regions where in situ measurements are sparse or unavailable. For these reasons, we focus our analysis on satellite-based datasets and identify where agreement between datasets exists.

Our results are robust (Extended Data Fig. 3) to a range of methodological assumptions including the length of analysis period, the choice of start and end period and the spatial extent of control pixels (Methods). Our analysis period includes the 2015–2016 El Niño that resulted in negative precipitation anomalies over many tropical land regions (Supplementary Fig. 1). We found that the precipitation response to forest loss was robustly negative during both El Niño and non-El Niño years (Extended Data Fig. 3). Over the Amazon and SEA, we see a stronger reduction in precipitation over regions of forest loss during El Niño years. The relative impact of El Niño on precipitation is smaller in the Congo[22], and correspondingly we do not see a stronger reduction here. A stronger precipitation response to forest loss in regions and periods impacted by El Niño is probably due to higher transpiration rates observed in tropical forests during El Niño years[23] and because rainfall is more sensitive to reductions in moisture recycling during drought years[5,24]. Climate change is expected to lead to increased droughts over many tropical regions[25], which may be further exacerbated by ongoing deforestation.

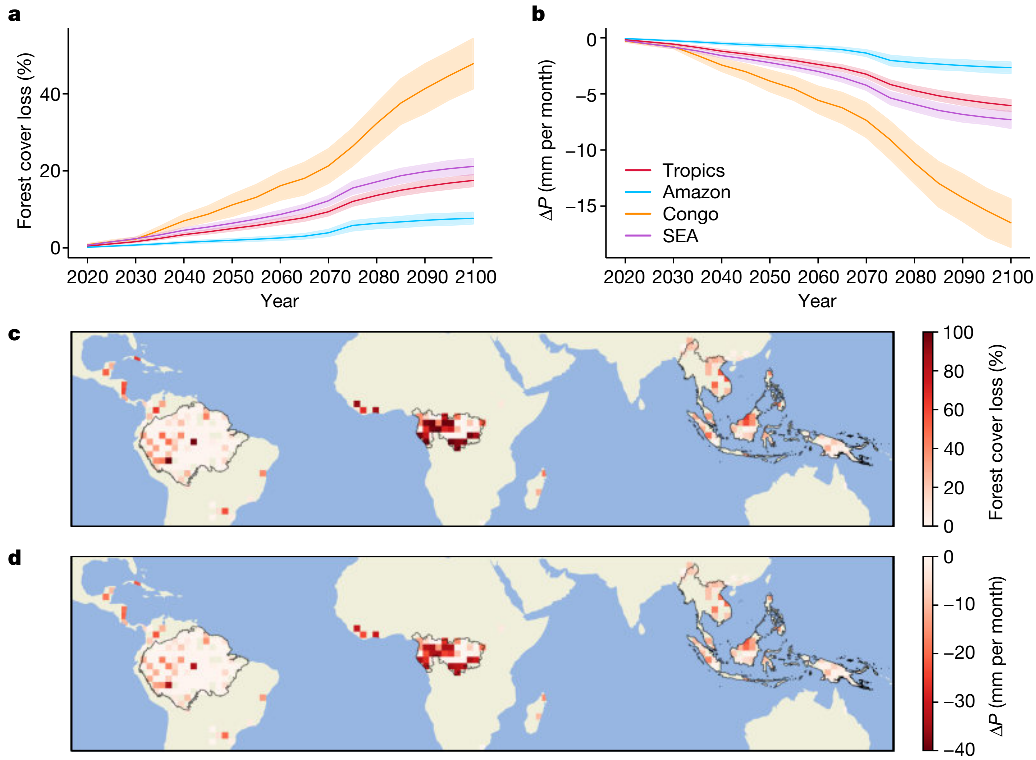


Fig. 4 | Impact of projected future forest loss on annual mean precipitation. **a**, Mean forest cover loss over 2015–2100 under Shared Socioeconomic Pathway 3–Representative Concentration Pathway 4.5 for the tropics, Amazon, Congo and SEA. **b**, Impact of projected forest cover loss on precipitation (*P*; ±1 standard error from the mean). **c**, Spatial pattern of forest cover loss. **d**, Predicted *P* change (Δ*P*) in 2100 due to forest cover loss. Results are shown for 2.0° resolution. Maps of the different regions generated using Cartopy and Natural Earth[51].

Seasonal precipitation reductions

Changes in precipitation due to forest loss during the dry, wet and transition seasons are nearly consistently negative (Fig. 3). For the tropics, absolute changes in precipitation with forest loss are greatest in the wet season (Fig. 3a, up to −0.6 mm per month per percentage point forest loss) whereas relative changes of precipitation with forest loss are similar (−0.2% per percentage point) across dry, wet and transition seasons (Supplementary Fig. 2). In the Amazon, deforestation causes the largest reductions in precipitation during the transition season (Fig. 3b) as has been found previously[18,26,27].

Previous case studies have indicated that dry-season precipitation can increase over deforestation in the Amazon[11,28,29]. We observed a nonsignificant increase in dry-season precipitation due to forest loss in the Amazon at 2° as well as increases in the Congo at 1° and 2° (Fig. 3). In SEA, forest loss causes reductions in dry-season precipitation across all scales (Fig. 3d). The mechanism through which forest loss impacts precipitation is likely to change with both season and spatial scale. At the smallest scales (5 km), thermally driven impacts are likely to dominate, shifting to dynamically driven impacts through reductions to surface roughness, then to reductions in moisture fluxes and precipitation recycling at the largest scales[12,30]. Our observation of greater reductions in precipitation due to deforestation at larger spatial scales is consistent with a reduction in moisture recycling emerging as the dominant mechanism[1].

Comparison with climate models

A meta-analysis of climate model studies (predominantly global models with >2° resolution) found that forest loss in the Amazon resulted in a mean reduction in annual mean precipitation of 0.16 ± 0.13% per percentage point[17], overlapping with our value of 0.25% per percentage point (Supplementary Fig. 2). Fewer simulations have been conducted for the Congo, with models predicting a reduction in precipitation of 0.16 ± 0.17% per percentage point[2], similar to our reduction of 0.15% per percentage point (Supplementary Fig. 2). The large range of model estimates highlights the substantial uncertainty in model predictions. Our observationally derived analysis provides support for models that predict reductions in precipitation under regional deforestation at global climate model scales.

Our observational analysis documents the impacts of deforestation on precipitation across the tropics. Applying linear scaling to the reductions in precipitation observed in our analysis would suggest that complete deforestation could result in reductions in annual precipitation of 10–20%. Previous estimates of the impact of complete deforestation on precipitation range from a 16% (ref. [17]) to 55–70% (ref. [31]) reduction in the Amazon and an 18% (ref. [2]) to 50% (ref. [32]) reduction in the Congo.

Impacts of future deforestation

To further explore how future deforestation might modify precipitation, we combined our observationally derived estimates of precipitation responses to forest cover loss with future projections of land cover change from a high-deforestation scenario (Methods). We estimate that forest loss from 2015 to 2100 (Fig. 4a) could lead to reductions of annual mean precipitation of up to 16.5 ± 6.2 mm per month in the Congo (Fig. 4b), equivalent to precipitation declines of 8–10%. Forest loss is projected to be greatest in the western and southern Congo (Fig. 4c), which will also experience the strongest reductions in precipitation (Fig. 4d).

The sensitivity of precipitation to the extent of forest loss is an uncertainty in our analysis, a result of the relatively short observational record, compounded by large spatial and temporal variability in precipitation. The response of precipitation to forest loss greater than

30%, a threshold beyond which large reductions in precipitation have been postulated[1], is one such uncertainty. Restricting our analysis to the forest losses of 0–30% that are well sampled in our observational dataset (Supplementary Fig. 3), through capping the impacts of greater forest loss at that of 30%, results in projected annual mean precipitation reductions of 6.5 ± 2.6 mm per month in the Congo and 6.2 ± 2.5 mm per month in SEA (Supplementary Fig. 4). However, restricting our analysis in this way is likely to underestimate the precipitation impacts over regions projected to experience the most extensive deforestation, including the Congo, where mean forest cover is projected to decline by 40 percentage points between 2015 and 2100 (Fig. 4a).

Previous studies have identified both linear[9,33] and nonlinear[1,31] responses of precipitation to forest loss. Such nonlinear interactions and feedbacks have the potential to further amplify or moderate the responses predicted here[14,34]. Our analysis shows large reductions in precipitation for relatively small amounts of forest loss and evidence for reduced sensitivity of precipitation to additional amounts of forest loss (Extended Data Fig. 1). Assuming a nonlinear relationship between forest loss and precipitation (Methods) reduces our projected reductions in precipitation by around a factor of 2 (Supplementary Fig. 5). Our observationally based approach will miss tipping points in the climate system that might be reached as deforestation extent progresses further[1]. Such tipping points have been postulated for the Amazon under future global change[25,35]. Thus, the substantial declines in precipitation projected in our analysis should be viewed as a conservative estimate of the potential precipitation response to future deforestation. Nevertheless, our analysis suggests that deforestation can drive local and regional precipitation changes that may match or exceed those predicted due to climate change over the same period[36,37].

Implications of precipitation reductions

Reductions in precipitation induced by forest loss have important implications for society and the sustainability of remaining tropical forest. Deforestation-induced reductions in precipitation affect agriculture[1,14] and hydropower generation[38]. On average, crop yields decline by 0.5% for each percentage point reduction in precipitation[39]. Our results indicate that forest-loss-induced changes to annual precipitation (Supplementary Fig. 2) could cause crop yields to decline by 1.25% for each 10-percentage-point loss of forest cover, potentially exacerbating the impacts of climate change and future drought events. The maintenance of regional rainfall patterns due to forests in the Amazon has been valued at up to US\$9 $ha^{-1}\,yr^{-1}$ and US\$1.84 $ha^{-1}\,yr^{-1}$ through sustaining agricultural yields and hydropower generation, respectively[40]. Global cropland area increased by 9% in the past two decades, with even higher increases in South America and tropical Africa[41] largely at the expense of natural ecosystems. Further agricultural expansion in tropical forest regions may lead to overall reductions in production if declines in yield due to deforestation-induced reductions in rainfall outweigh increased production from expanded agricultural area[14].

Furthermore, reductions in rainfall over remaining areas of tropical forest are expected to lead to additional forest loss[9] as well as impacting species composition[22], carbon sequestration[42] and fire frequency[43]. Reductions in dry-season precipitation pose a particular threat to forest viability by exacerbating seasonal droughts and potentially delaying the onset of the wet season and extending the length of the dry season. Increases in dry-season length over recent decades have previously been reported for the Amazon[44] and the Congo[45], possibly linked to land cover changes[27].

Deforestation may also shift precipitation patterns, increasing dry-season rainfall immediately downwind of forest loss and decreasing rainfall in upwind areas[12]. Our approach is restricted to observing deforestation impacts up to scales of 200 km (Methods). At larger scales, insufficient pixels experienced forest loss during the relatively short period of satellite observations for a robust analysis. Deforestation is also likely to alter precipitation at these larger scales through reducing moisture recycling leading to reductions in rainfall downwind of forest loss[4,5,9,35]. The length scale of moisture recycling has been estimated at 500–2,000 km in the tropics[46], with a median value of 600 km in the Amazon[5]. In regions downwind of extensive forests, such as the south-western Amazon, up to 70% of precipitation could be sourced from upwind evapotranspiration[47,48]. Tropical forest loss could therefore have severe implications for precipitation in these regions that are hundreds to thousands of kilometres downwind of the forest loss[5]. Through missing the impacts at these larger scales, our analysis is likely to underestimate the full impacts of deforestation on rainfall.

Our results highlight the importance of remaining tropical forests for sustaining regional precipitation. Despite efforts to reduce deforestation, rates of tropical forest loss have accelerated over the past two decades[49]. Renewed efforts are needed to ensure recent commitments to reduce deforestation, including the New York Declaration on Forests and The Glasgow Leaders' Declaration on Forests and Land Use made at the 26th UN Climate Change Conference of the Parties, are successful. Global efforts to restore large areas of degraded and deforested land could enhance precipitation[50], reversing some of the reductions in precipitation due to forest loss observed here.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-05690-1.

1. Lawrence, D. & Vandecar, K. Effects of tropical deforestation on climate and agriculture. *Nat. Clim. Change* **5**, 27–36 (2015).
2. Spracklen, D. V., Baker, J. C. A., Garcia-Carreras, L. & Marsham, J. H. The effects of tropical vegetation on rainfall. *Annu. Rev. Environ. Resour.* **43**, 193–218 (2018).
3. Bonan, G. B. Forests and climate change: forcings, feedbacks, and the climate benefits of forests. *Science* **320**, 1444–1449 (2008).
4. Spracklen, D. V., Arnold, S. R. & Taylor, C. M. Observations of increased tropical rainfall preceded by air passage over forests. *Nature* **489**, 282–285 (2012).
5. Staal, A. et al. Forest-rainfall cascades buffer against drought across the Amazon. *Nat. Clim. Change* **8**, 539–543 (2018).
6. Baker, J. C. A. & Spracklen, D. V. Divergent representation of precipitation recycling in the Amazon and the Congo in CMIP6 models. *Geophys. Res. Lett.* **49**, e2021GL095136 (2022).
7. Guan, K. et al. Photosynthetic seasonality of global tropical forests constrained by hydroclimate. *Nat. Geosci.* **8**, 284–289 (2015).
8. Staal, A. et al. Hysteresis of tropical forests in the 21st century. *Nat. Commun.* **11**, 4978 (2020).
9. Zemp, D. C. et al. Self-amplified Amazon forest loss due to vegetation-atmosphere feedbacks. *Nat. Commun.* **8**, 14681 (2017).
10. Hansen, M. C. et al. High-resolution global maps of 21st-century forest cover change. *Science* **342**, 850–854 (2013).
11. Chagnon, F. J. F. & Bras, R. L. Contemporary climate change in the Amazon. *Geophys. Res. Lett.* **32**, L13703 (2005).
12. Khanna, J., Medvigy, D., Fueglistaler, S. & Walko, R. Regional dry-season climate changes due to three decades of Amazonian deforestation. *Nat. Clim. Change* **7**, 200–204 (2017).
13. Garcia-Carreras, L. & Parker, D. J. How does local tropical deforestation affect rainfall? *Geophys. Res. Lett.* **38**, L19802 (2011).
14. Leite-Filho, A. T., Soares-Filho, B. S., Davis, J. L., Abrahão, G. M. & Börner, J. Deforestation reduces rainfall and agricultural revenues in the Brazilian Amazon. *Nat. Commun.* **12**, 2591 (2021).
15. McAlpine, C. A. et al. Forest loss and Borneo's climate. *Environ. Res. Lett.* **13**, 044009 (2018).
16. Chapman, S. et al. Compounding impact of deforestation on Borneo's climate during El Niño events. *Environ. Res. Lett.* **15**, 084006 (2020).
17. Spracklen, D. V. & Garcia-Carreras, L. The impact of Amazonian deforestation on Amazon basin rainfall. *Geophys. Res. Lett.* **42**, 9546–9552 (2015).
18. Jiang, Y. et al. Modeled response of South American climate to three decades of deforestation. *J. Clim.* **34**, 2189–2203 (2021).
19. Harris, I., Osborn, T. J., Jones, P. & Lister, D. Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Sci. Data* **7**, 109 (2020).
20. Fassoni-Andrade, A. C. et al. Amazon hydrology from space: scientific advances and future challenges. *Rev. Geophys.* **59**, e2020RG000728 (2021).
21. Haiden, T., Janousek, M., Vitart, F., Ferranti, L. & Prates, F. *Evaluation of ECMWF Forecasts, Including the 2019 Upgrade. ECMWF Technical Memorandum No. 853* (ECMWF, 2019).
22. Esquivel-Muelbert, A. et al. Compositional response of Amazon forests to climate change. *Glob. Change Biol.* **25**, 39–56 (2019).

23. Brum, M. et al. ENSO effects on the transpiration of eastern Amazon trees. *Philos. Trans. R. Soc. B* **373**, 20180085 (2018).
24. Bagley, J. E., Desai, A. R., Harding, K. J., Snyder, P. K. & Foley, J. A. Drought and deforestation: has land cover change influenced recent precipitation extremes in the Amazon? *J. Clim.* **27**, 345–361 (2014).
25. Wunderling, N. et al. Recurrent droughts increase risk of cascading tipping events by outpacing adaptive capacities in the Amazon rainforest. *Proc. Natl Acad. Sci. USA* **119**, e2120777119 (2022).
26. Fu, R. & Li, W. The influence of the land surface on the transition from dry to wet season in Amazonia. *Theor. Appl. Climatol.* **78**, 97–110 (2004).
27. Leite-Filho, A. T., de Sousa Pontes, V. Y. & Costa, M. H. Effects of deforestation on the onset of the rainy season and the duration of dry spells in southern Amazonia. *J. Geophys. Res. Atmos.* **124**, 5268–5281 (2019).
28. Negri, A. J., Adler, R. F., Xu, L. & Surratt, J. The Impact of Amazonian deforestation on dry season rainfall. *J. Clim.* **17**, 1306–1319 (2004).
29. Chagnon, F. J. F., Bras, R. L. & Wang, J. Climatic shift in patterns of shallow clouds over the Amazon. *Geophys. Res. Lett.* **31**, L24212 (2004).
30. Chambers, J. Q. & Artaxo, P. Biosphere–atmosphere interactions: deforestation size influences rainfall. *Nat. Clim. Change* **7**, 175–176 (2017).
31. Baudena, M., Tuinenburg, O. A., Ferdinand, P. A. & Staal, A. Effects of land-use change in the Amazon on precipitation are likely underestimated. *Glob. Change Biol.* **27**, 5580–5587 (2021).
32. Duku, C. & Hein, L. The impact of deforestation on rainfall in Africa: a data-driven assessment. *Environ. Res. Lett.* **16**, 064044 (2021).
33. Akkermans, T., Thiery, W. & Van Lipzig, N. P. M. The regional climate impact of a realistic future deforestation scenario in the Congo basin. *J. Clim.* **27**, 2714–2734 (2014).
34. Staal, A. et al. Feedback between drought and deforestation in the Amazon. *Environ. Res. Lett.* **15**, 044024 (2020).
35. Xu, X. et al. Deforestation triggering irreversible transition in Amazon hydrological cycle. *Environ. Res. Lett.* **17**, 034037 (2022).
36. Kooperman, G. J. et al. Forest response to rising CO_2 drives zonally asymmetric rainfall change over tropical land. *Nat. Clim. Change* **8**, 434–440 (2018).
37. Chen, Z. et al. Global land monsoon precipitation changes in CMIP6 projections. *Geophys. Res. Lett.* **47**, e2019GL086902 (2020).
38. Stickler, C. M. et al. Dependence of hydropower energy generation on forests in the Amazon Basin at local and regional scales. *Proc. Natl Acad. Sci. USA* **110**, 9601–9606 (2013).
39. Challinor, A. J. et al. A meta-analysis of crop yield under climate change and adaptation. *Nat. Clim. Change* **4**, 287–291 (2014).
40. Strand, J. et al. Spatially explicit valuation of the Brazilian Amazon forest's ecosystem services. *Nat. Sustain.* **1**, 657–664 (2018).
41. Potapov, P. et al. Global maps of cropland extent and change show accelerated cropland expansion in the twenty-first century. *Nat. Food* **3**, 19–28 (2022).
42. Li, Y. et al. Deforestation-induced climate change reduces carbon storage in remaining tropical forests. *Nat. Commun.* **13**, 1964 (2022).
43. Aragão, L. E. O. C. et al. Interactions between rainfall, deforestation and fires during recent years in the Brazilian Amazonia. *Philos. Trans. R. Soc. B* **363**, 1779–1785 (2008).
44. Marengo, J. A. et al. Changes in climate and land use over the Amazon region: current and future variability and trends. *Front. Earth Sci.* https://doi.org/10.3389/feart.2018.00228 (2018).
45. Jiang, Y. et al. Widespread increase of boreal summer dry season length over the Congo rainforest. *Nat. Clim. Change* https://doi.org/10.1038/s41558-019-0512-y (2019).
46. Van Der Ent, R. J. & Savenije, H. H. G. Length and time scales of atmospheric moisture recycling. *Atmos. Chem. Phys.* **11**, 1853–1863 (2011).
47. Sorí, R., Nieto, R., Vicente-Serrano, S. M., Drumond, A. & Gimeno, L. A Lagrangian perspective of the hydrological cycle in the Congo River basin. *Earth Syst. Dyn.* **8**, 653–675 (2017).
48. van der Ent, R. J., Savenije, H. H. G., Schaefli, B. & Steele-Dunne, S. C. Origin and fate of atmospheric moisture over continents. *Water Resour. Res.* **46**, W09525 (2010).
49. Feng, Y. et al. Doubling of annual forest carbon loss over the tropics during the early twenty-first century. *Nat. Sustain.* **4**, 441–451 (2022).
50. Tuinenburg, O. A., Bosmans, J. H. C. & Staal, A. The global potential of forest restoration for drought mitigation. *Environ. Res. Lett.* **17**, 034045 (2022).
51. Met Office. Cartopy: a cartographic python library with a Matplotlib interface 2010–2015. *Met Office* https://scitools.org.uk/cartopy (2022).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

Methods

Datasets

We used 18 precipitation datasets, listed in Extended Data Table 1. All datasets were downloaded at the highest available spatial resolution, which for some datasets was 0.04°, or approximately 4 km at the Equator. Data were obtained as monthly means or converted to monthly mean using the Python package xarray[52]. We categorized precipitation datasets as satellite ($n = 10$), station ($n = 4$) and reanalysis ($n = 4$). Satellite datasets are those based primarily on data from satellite sensors and include datasets that have both satellite and station-based data (that is, merged datasets). Station datasets include only ground-based information from weather stations and rain gauges. Reanalysis products are models constrained by surface and satellite data. Precipitation datasets have been compared previously over the Amazon[20] highlighting the limited station data over tropical forest regions. Time series of precipitation (Supplementary Fig. 1) reveal variability across the different datasets highlighting the need to analyse impacts of deforestation across multiple datasets.

To analyse the changes in forest canopy cover, we used data from the Global Forest Change (GFC) version 1.9 (ref. [10]). GFC v1.9 provides forest canopy cover in the year 2000 and subsequent annual forest loss from 2001–2020 at 30-m resolution. We analysed forest cover and precipitation changes over the period 2003 to 2017, which was the period common to all datasets.

Analysis across multiple spatial scales

We analysed the impacts of forest loss across a range of scales (0.05°, 0.1°, 0.25°, 0.5°, 1.0° and 2.0°). Each precipitation dataset was analysed at its native resolution and at all lower resolutions across this range of scales. Spatial regridding was carried out using the Python package xESMF[53] with a bilinear regridding scheme. Two alternative regridding methods (xESMF: conservative-normalized; and iris: area weighted) were tested and had little impact on our results. For GFC data, we calculated forest loss using the original 30-m data and converted the resulting values to each of the six spatial resolutions analysed by taking the sum of all 30-m pixels within each larger pixel. Change in canopy cover from 2003 to 2017 at each resolution is shown in Fig. 1.

Assessing impact of historical deforestation on precipitation

We used a moving-window nearest-neighbour approach[54] to compare the forest loss and precipitation change of each pixel with that of its immediate neighbours. We tested the sensitivity of the analysis to the size of the moving window and found similar results for 3 × 3 and 5 × 5 (Extended Data Fig. 2) moving windows. Results from the 3 × 3 moving-window approach can been seen in the main paper. We calculated the forest loss of each deforested pixel relative to neighbouring control pixels as the forest loss of the deforested pixel minus forest loss of the control. We constrained our analysis to the tropical evergreen broadleaf biome using the Moderate Resolution Imaging Spectroradiometer land cover dataset[55]. To be included in the analysis, deforested pixels must have experienced 0.1% more forest loss over time than their neighbouring control pixels. The number of deforested pixels analysed varied between analysis resolutions as follows: 0.05°, $n = 243{,}254$; 0.1°, $n = 58{,}660$; 0.25°, $n = 9{,}604$; 0.5°, $n = 2{,}303$; 1.0°, $n = 586$; 2.0°, $n = 123$. We observed similar distributions of canopy change for all spatial resolutions analysed (Supplementary Fig. 6).

We calculated the precipitation change of the deforested pixel relative to the precipitation change of the control pixel (ΔP) as the precipitation change of the deforested pixel over the analysis period (for example, 2003–2017) minus the precipitation change over the control pixel. To reduce the impact of interannual variability in precipitation on our results, we calculated 5-yr means for periods at the start (2003–2007) and end (2013–2017; Extended Data Fig. 5) of the analysis period. We then calculated the change in precipitation as the difference between the start and end of these multi-year means. We report precipitation changes (ΔP) as a function of forest loss by dividing by the difference in forest loss between deforestation and control pixels (units of millimetres per month per percentage point). We also report precipitation change as the percentage change in precipitation ($\Delta P/P$, in units of per cent) as a function of forest loss (in units of per cent per percentage point).

To ensure that control pixels and deforested pixels experience a similar background climate, we conducted a sensitivity test in which we restricted our analysis to pixels for which the pre-deforestation precipitation across the control and deforested pixels differed by less than 10%. Restricting our analysis in this way had little impact on our results (Supplementary Fig. 7) showing that our nearest-neighbour approach is effective even at the largest scales analysed here.

To explore the role of the analysis period on our results, we compared the results for 5-yr means to those for shorter 3-yr means (2003–2005 versus 2015–2017) and found consistent results (Extended Data Fig. 3). Our analysis period includes the strong 2015/2016 El Niño that resulted in reductions in precipitation over most tropical land regions, particularly in 2015 (Supplementary Fig. 1). To explore the potential impacts of the 2015/2016 El Niño on our analysis, we estimated the impact of forest loss on precipitation using 3-yr (2003–2005 versus 2018–2020) and 5-yr (2003–2007 versus 2016–2020) multi-annual means spanning an extended time period. The 3-yr analysis completely excludes the 2015/2016 ENSO, and the 5-yr analysis excludes 2015, which was the driest year (Extended Data Fig. 3). Two datasets (TRMM and UDEL) were not available after 2017, so they were removed from this sensitivity analysis.

Statistical analysis

For each category of precipitation data (satellite, station and reanalysis), precipitation change values were grouped together for all deforestation pixels and all control pixels. We found that precipitation changes for deforested pixels and control pixels, and the difference in precipitation change between deforested and control pixels (Extended Data Fig. 4), were normally distributed. Error bars (Figs. 2 and 3) show ±1 standard error from the mean calculated and displayed using the Python package Seaborn[56]. To test whether mean precipitation changes over regions of deforestation were statistically different from changes over the control areas, we used a Student's t-test. We also used the Mann–Whitney test to test for significant differences in median precipitation change between control and deforested pixels and found similar results.

Seasonal analysis

For the satellite datasets alone, in addition to calculating precipitation changes at the annual timescale, we calculated changes for the dry season (driest 3 months of each year), wet season (wettest 3 months of each year) and transition season (remaining 6 months). The driest, wettest and transition months were identified for each pixel using each individual precipitation dataset. For each season and dataset, we calculated the median change in precipitation across all of the pixels within the region of interest (Supplementary Figs. 8–10).

Predicting future precipitation change due to forest loss

We used projections of forest cover change available at 0.05° from the Global Change Analysis Model (GCAM) for 2015–2100 based on the Shared Socioeconomic Pathway 3–Representative Concentration Pathway 4.5 scenario, which represents a high-deforestation future[57]. GCAM includes the impacts of climate and land use on future forest cover. We summed forest cover from all forest categories and calculated forest cover loss in each year compared to a 2015 baseline. Forest cover loss data were regridded to 2°. We estimated the impact of forest loss on future precipitation at the 2° scale through multiplying the projected percentage point forest loss for each pixel by the

observed median change in precipitation (millimetres per month) per percentage point forest cover loss across the satellite datasets. To estimate the uncertainty in our predictions, we applied an upper and lower limit on the sensitivity of precipitation to forest loss based on the median value ±1 standard error from the mean (see error bars in Fig. 2) and rescaled by forest loss. This provides a range of estimated precipitation impacts of future forest loss. We also tested the impact on our results of capping future forest loss in each pixel at 30%, which is the upper range of forest loss that is well sampled in the observations (Supplementary Fig. 3). For each region, we applied the tropical satellite precipitation response to forest loss (Fig. 2f), meaning that our projected regional precipitation changes are a product of the regional canopy cover change and the median tropical precipitation response. Our approach assumes a linear precipitation response to forest loss, which recent work suggests could provide a conservative estimate of deforestation impacts[31]. We tested the sensitivity of assuming a linear response of precipitation to canopy cover loss. We fitted a nonlinear function to the data presented in Extended Data Fig. 1 through applying the median sensitivity of precipitation to forest cover loss (millimetres per month per percentage point) within each forest cover loss bin. We then scaled by the projected forest cover loss. This approach reduces the projected reduction in precipitation to 2.4 mm per month in SEA and 1.5 mm per month in the Congo (Supplementary Fig. 5).

Data availability

Full results for all tested resolutions used in this analysis are available through https://doi.org/10.5281/zenodo.7373832. The original datasets are freely available to download from the following repositories: CHIRPS from https://data.chc.ucsb.edu/products/?C=M;O=D, CMORPH from https://ftp.cpc.ncep.noaa.gov/precip/CMORPH_RT/GLOBE/data/, CPC from https://psl.noaa.gov/data/gridded/data.cpc.globalprecip.html, CRU from https://crudata.uea.ac.uk/cru/data/hrg/, ERA5 from https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview, GPCC from https://opendata.dwd.de/climate_environment/GPCC/html/download_gate.html, GPCP from https://disc.gsfc.nasa.gov/datasets/GPCPMON_3.1/summary?keywords=GPCPMON, GPM from https://gpm1.gesdisc.eosdis.nasa.gov/data/GPM_L3/, JRA from https://climatedataguide.ucar.edu/climate-data/jra-55 and https://jra.kishou.go.jp/JRA-55/index_en.html, MERRA-2 from https://disc.gsfc.nasa.gov/datasets?project=MERRA-2, NOAA (PREC/LAND) from https://psl.noaa.gov/data/gridded/data.precl.html, PERSIANN (CCS, CDR, CCS-CDR, PDIR-NOW) from https://chrsdata.eng.uci.edu/, TRMM from https://disc.gsfc.nasa.gov/datasets/TRMM_3B43_7/summary, UDEL from https://psl.noaa.gov/data/gridded/data.UDel_AirT_Precip.html. The GCAM model output used in this study is available from https://doi.org/10.25584/data.2020-07.1357/1644253. Source data are provided with this paper.

Code availability

The code used in this analysis is available through https://doi.org/10.5281/zenodo.7373832.

52. Hoyer, S. & Hamman, J. xarray: N-D labeled arrays and datasets in Python. *J. Open Res. Softw.* https://doi.org/10.5334/jors.148 (2017).
53. Zhuang, J. xESMF. *Zenodo* https://doi.org/10.5281/zenodo.1134365 (2022).
54. Baker, J. C. A. & Spracklen, D. V. Climate benefits of intact Amazon forests and the biophysical consequences of disturbance. *Front. For. Glob. Change* https://doi.org/10.3389/ffgc.2019.00047 (2019).
55. Schaaf, C. & Wang, Z. MCD43A3 MODIS/Terra+Aqua BRDF/Albedo Daily L3 Global - 500m V006. *NASA EOSDIS Land Processes DAAC* https://doi.org/10.5067/modis/mcd43a3.006 (2015).
56. Waskom, M. Seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
57. Chen, M. et al. Global land use for 2015–2100 at 0.05° resolution under diverse socioeconomic and climate scenarios. *Sci. Data* **7**, 320 (2020).
58. Funk, C. et al. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Sci. Data* **2**, 150066 (2015).
59. Xie, P. et al. NOAA Climate Data Record (CDR) of CPC Morphing technique (CMORPH) high resolution global precipitation estimates, version 1. *NOAA National Centers for Environmental Information* https://doi.org/10.25921/w9va-q159 (2019).
60. Xie, P. et al. A gauge-based analysis of daily precipitation over East Asia. *J. Hydrometeorol.* **8**, 607–626 (2007).
61. Hersbach, H. et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).
62. Elke, R., Hänsel, S., Finger, P., Schneider, U. & Ziese, M. GPCC Climatology Version 2022 at 0.25°: monthly land-surface precipitation climatology for every month and the total year from rain-gauges built on GTS-based and historical data. *GPCC* https://doi.org/10.5676/DWD_GPCC/CLIM_M_V2022_025 (2022).
63. Huffman, G. J. A., Behrangi, R. F., Adler, D. T., Bolvin, E. J. & Nelkin, G. G. Introduction to the new version 3 GPCP monthly global precipitation analysis. *GPCP* https://docserver.gesdisc.eosdis.nasa.gov/public/project/MEaSUREs/GPCP/Release_Notes.GPCPV3.2.pdf (2022).
64. Hou, A. Y. et al. The global precipitation measurement mission. *Bull. Am. Meteorol. Soc.* **95**, 701–722 (2014).
65. Kobayashi, S. et al. The JRA-55 reanalysis: general specifications and basic characteristics. *J. Meteorol. Soc. Japan* **93**, 5–48 (2015).
66. Gelaro, R. et al. The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *J. Clim.* **30**, 5419–5454 (2017).
67. Chen, M., Xie, P. & Janowiak, J. E. Global land precipitation: a 50-yr monthly analysis based on gauge observations. *J. Hydrometeorol.* **3**, 249–266 (2002).
68. Nguyen, P. et al. The CHRS data portal, an easily accessible public repository for PERSIANN global satellite precipitation data. *Sci. Data* **6**, 1180296 (2019).
69. Ashouri, H. et al. PERSIANN-CDR: daily precipitation climate data record from multisatellite observations for hydrological and climate studies. *Bull. Am. Meteorol. Soc.* **96**, 69–83 (2015).
70. Nguyen, P. et al. Persiann dynamic infrared–rain rate (PDIR-now): a near-real-time, quasi-global satellite precipitation dataset. *J. Hydrometeorol.* **21**, 2893–2906 (2020).
71. Sadeghi, M. et al. PERSIANN-CCS-CDR, a 3-hourly 0.04° global precipitation climate data record for heavy precipitation studies. *Sci. Data* **8**, 157 (2021).
72. Huffman, G. J. et al. The TRMM Multisatellite Precipitation Analysis (TMPA): quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *J. Hydrometeorol.* **8**, 38–55 (2007).
73. Matsuura, K. & Willmott, C. J. Terrestrial precipitation: 1900-2017 gridded monthly time series. *Global Precipitation Archive* http://climate.geog.udel.edu/~climate/html_pages/Global2017/README.GlobalTsP2017.html (2018).

Acknowledgements The research has been supported by funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (DECAF project, grant agreement no. 771492), and the Newton Fund, through the Met Office Climate Science for Service Partnership Brazil.

Author contributions All authors developed the concept of the study, and contributed to experimental design, interpretation of results and writing of the manuscript. C.S. and J.C.A.B. contributed to the analysis.

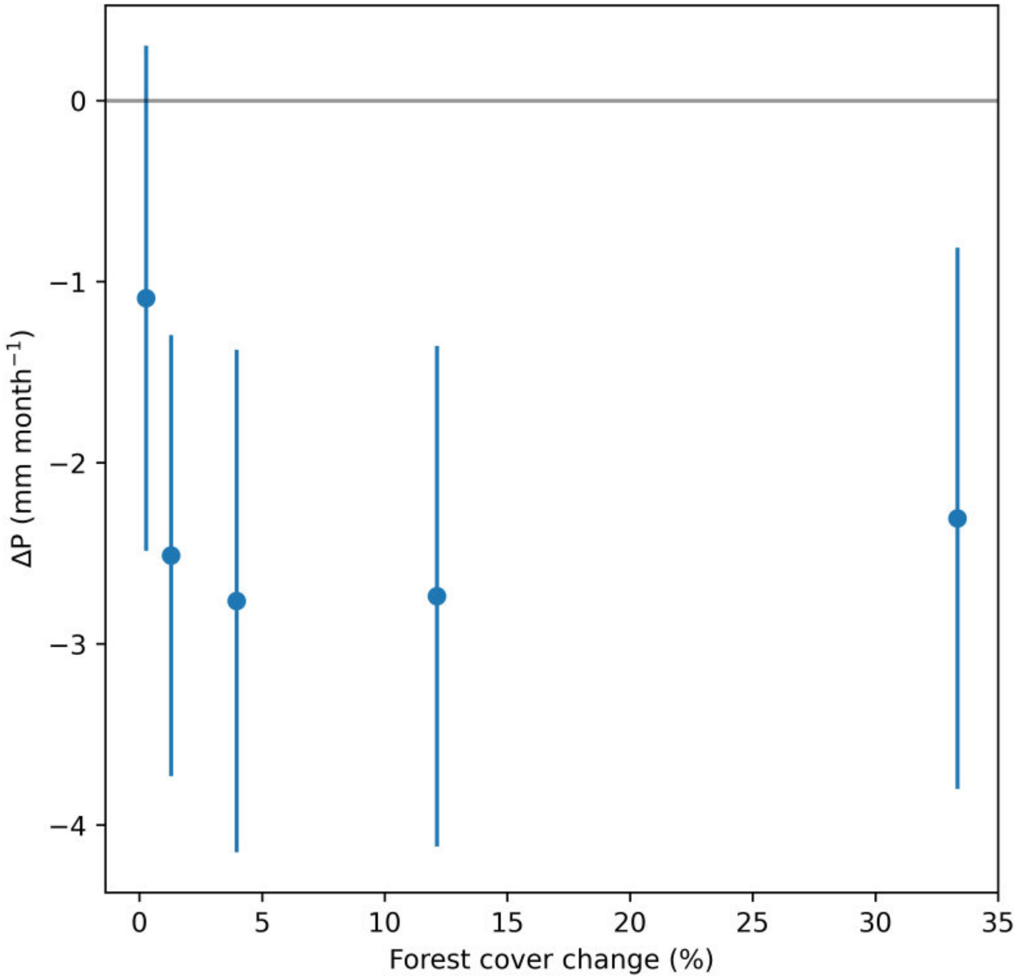
Competing interests The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-05690-1.

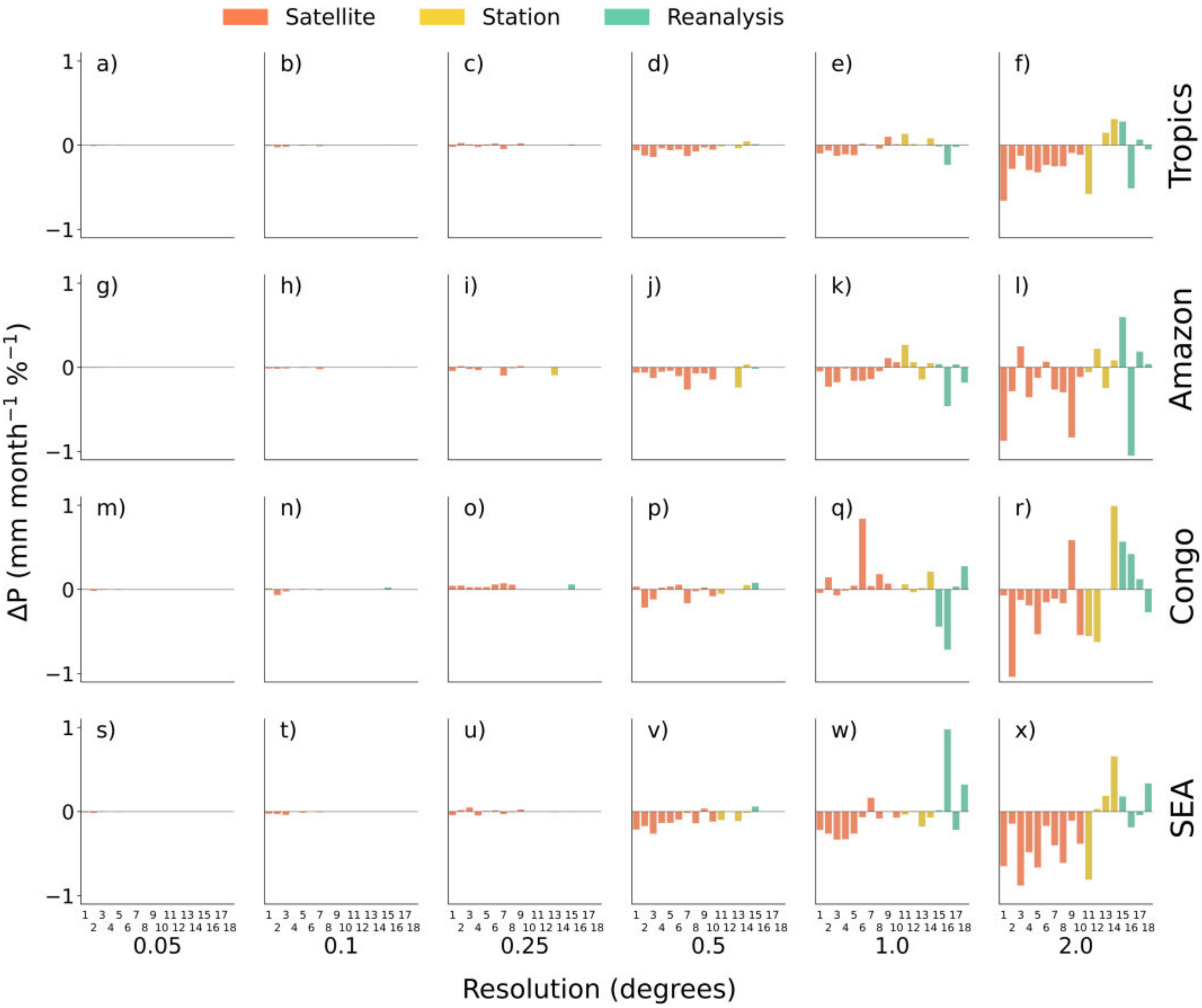
Correspondence and requests for materials should be addressed to C. Smith.

Peer review information *Nature* thanks Arie Staal and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

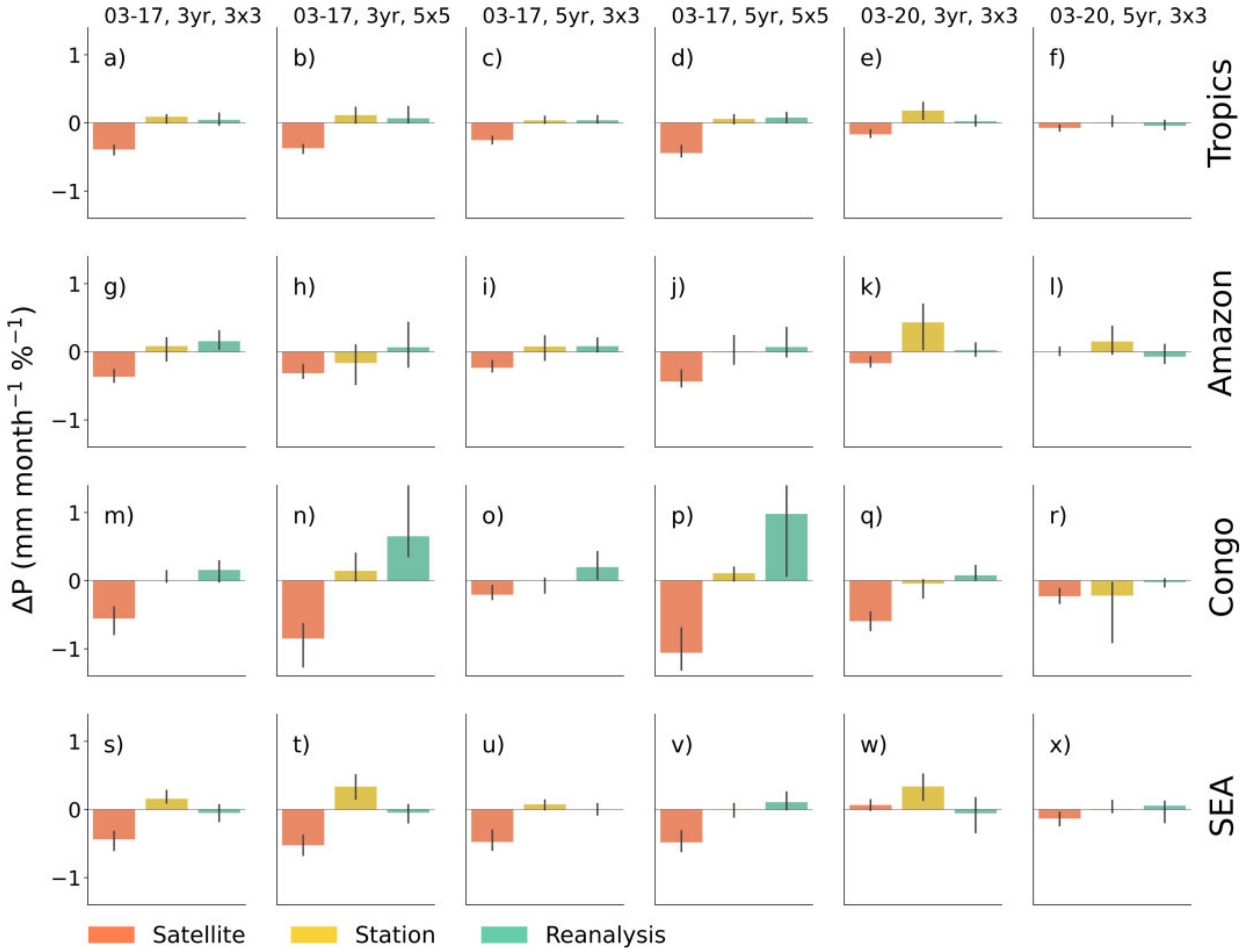
Reprints and permissions information is available at http://www.nature.com/reprints.



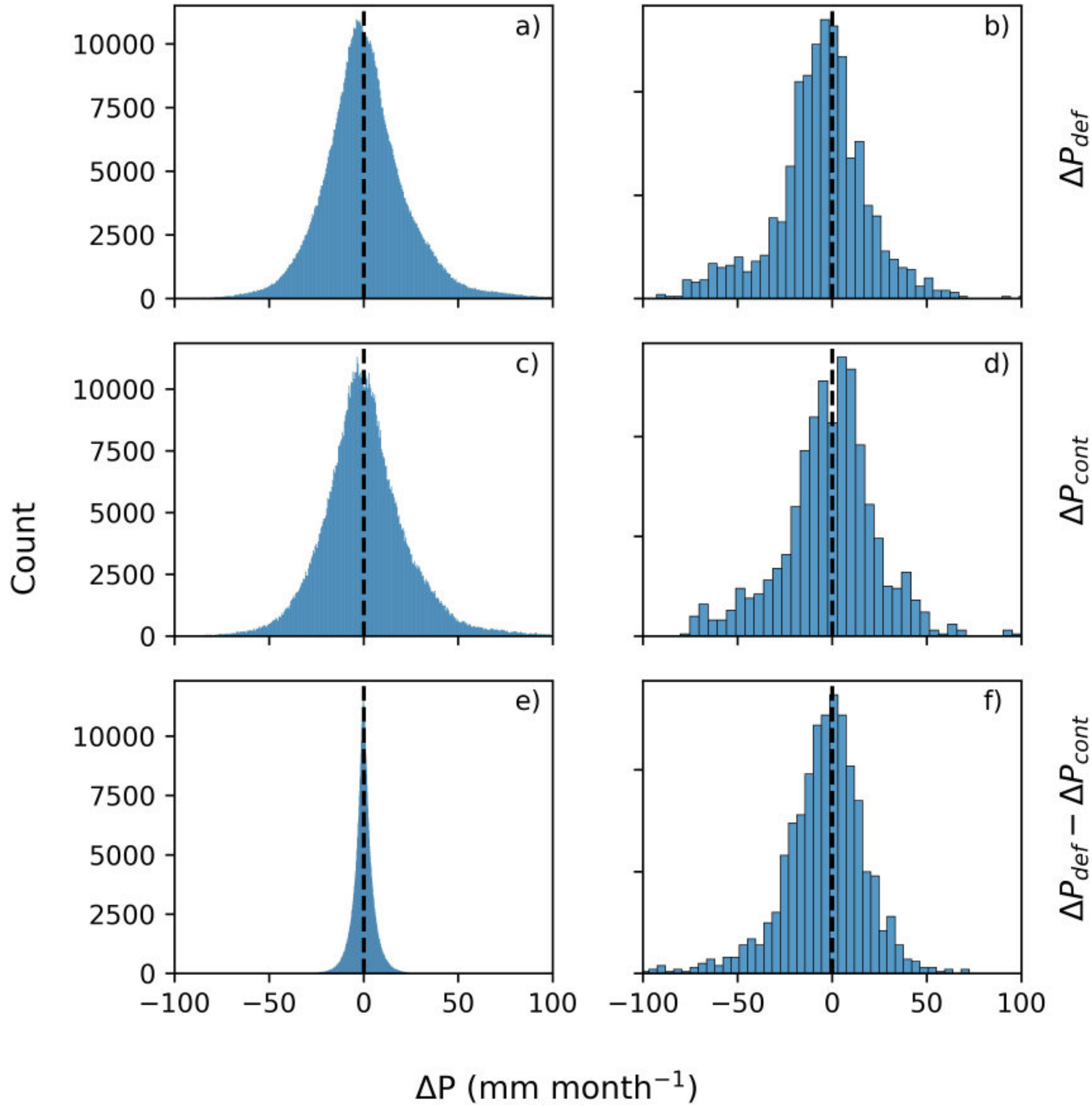
Extended Data Fig. 1 | Annual precipitation change as a function of forest loss. Results are shown at 2° spatial resolution for all satellite precipitation (P) datasets calculated as the change in P over time for deforested data pixels minus change over time for control data pixels. Data is binned according to forest cover change (%) with an equal number of pixels in each bin. Points show the median and error bars show ±1 standard error from the mean. Details of each data product are provided in Extended Data Table 1.



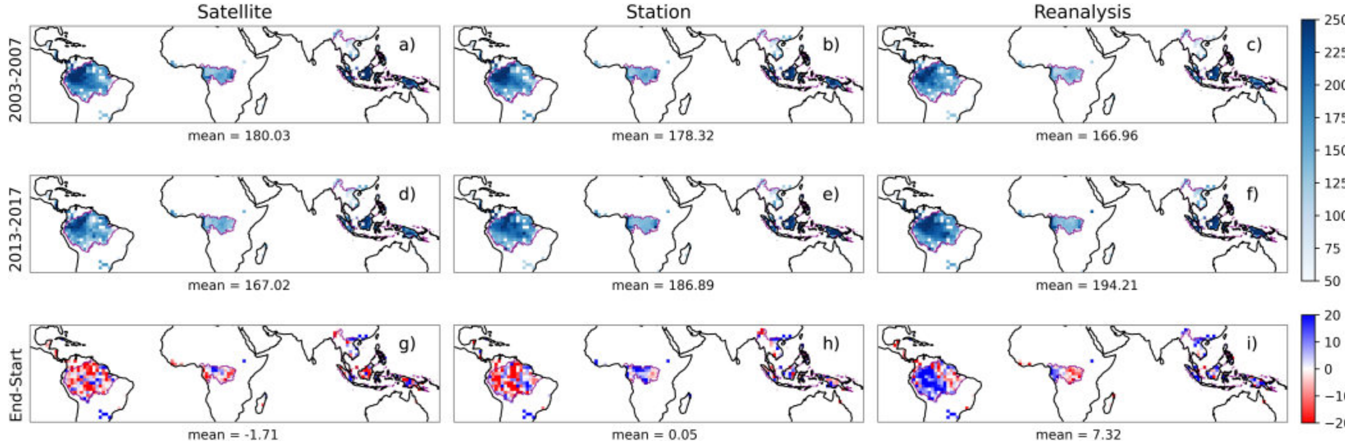
Extended Data Fig. 2 | Annual precipitation change due to forest loss for individual datasets. Results are shown for 2003 – 2017 for 5 year averages and 3x3 moving window. Bars show the median absolute change in annual P (mm month^{-1}) per percentage point tree cover loss in each region (Tropics (a-f), Amazon (g-l), Congo (m-r), SEA (s-x)). Each P dataset is shown separately and ordered and coloured by category: satellite (orange), station (yellow) and reanalysis (turquoise). The datasets are numbered; 1) CHIRPS, 2) CMORPH, 3) CPC, 4) CRU, 5) ERA5, 6) GPCC, 7) GPCP, 8) GPM, 9) JRA, 10) MERRA-2, 11) NOAA 12) PERSIANN-CCS, 13) PERSIANN-CCSCDR, 14) PERSIANN-CDR, 15) PERSIANN-NOW, 16) PERSIANN, 17) TRMM, 18) UDEL. Results are shown for forest loss scales of 0.05° (a,g,m,s), 0.1° (b,h,n,t), 0.25° (c,i,o,u), 0.5° (d,j,p,v), 1.0° (e,k,q,w), 2.0° (f,l,r,x). Details of each data product are provided in Extended Data Table 1.



Extended Data Fig. 3 | Changes in precipitation due to forest loss for different time periods and nearest neighbour comparisons. Changes in annual mean precipitation at 2.0° resolution are shown for satellite (orange), station (yellow) and reanalysis (turquoise) datasets for the tropics (a-f), Amazon (g-l), Congo (m-r) and Southeast Asia (SEA, s-x). Columns show the sensitivity of our results to changes in the analysis period, number of years used to compute multi-annual means at start and end of the analysis period, and size of the moving window used for nearest neighbour comparisons: 2003-2017, 3-year averages and 3x3 nearest neighbour (Column 1, a,g,m,s); 2003-2017, 3-year, 5x5 (Column 2; b,g,n,t); 2003-2017, 5-year, 3x3 (Column 3; c,i,o,u); 2003-2017, 5-year, 5x5 (Column 4; d,j,p,v); 2003-2020, 3-year, 3x3 (Column 5; e,k,q,w); 2003-2020, 5-year, 3x3 (Column 6; f,l,r,x). Error bars show ± 1 standard error from the mean. Details of each data product are provided in Extended Data Table 1. Full results for all tested resolutions are available in an online repository [10.5281/zenodo.7373832].



Extended Data Fig. 4 | Change in precipitation over deforested, control and difference between deforested and control pixels. Change in precipitation over 2003 to 2017 is shown for deforested (a, b), control (c, d) and difference between deforested and control pixels (e, f) for 0.05° (a, c, e) and 2.0° (b, d, f) resolution. Details of each data product are provided in Extended Data Table 1.



Extended Data Fig. 5 | Mean precipitation from satellite, station and reanalysis datasets. For each class of dataset, satellite (a, d, g), station (b, e, h) and reanalysis (c, f, i), the median value for the 5-year multi-annual mean at the start (2003-2007; a, b, c) and end (2013-2017; d, e, f) of the analysis period as well as the change over the analysis period (end – start; g, h, i) is shown. Mean values across tropical evergreen broadleaf forests are shown in units of mm/month at the top of each panel. Maps of the different regions generated using Cartopy and Natural Earth[51]. Details of each data product are provided in Extended Data Table 1.

Extended Data Table 1 | Precipitation datasets used in this study refs[58–73]

Dataset	Date Range	Highest Res	Inputs	Category	Reference
CHIRPS v2.0	1981 - 2021	0.05	Satellite + station	Satellite	Funk et al., 2015 [57]
CMORPH	1998 - 2020	0.25	Satellite + GPCP	Satellite	Xie et al., 2019 [58]
CPC	1979 - 2021	0.5	GTS + COOP	Station	Xie et al., 2007 [59]
CRU TS v4.06	1901 - 2021	0.5	Station	Station	Harris et al., 2020 [19]
ERA5	1979 - 2021	0.1	Numerous	Reanalysis	Hersbach et al., 2020 [60]
GPCC v2022	1891 - 2020	0.25	Stations	Station	Elke et al., 2022 [61]
GPCP v3.2	1996 -2020	0.5	PERSIANN, GPROF TORVS/AIRS, GPCC GAUGE	Satellite	Huffman et al., 2022 [62]
GPM v0.6	2000 - 2020	0.1	9 satellites	Satellite	Hou et al., 2014 [63]
JRA v7.0	1979 - 2020	0.5625	Numerous	Reanalysis	Kobayashi et al., 2015 [64]
MERRA-2	1980 - 2020	0.625	Numerous	Reanalysis	Gelaro et al., 2017 [65]
NOAA	1948 - 2021	1	GHCN gauge, CAMS	Reanalysis	Chen et al., 2002 [66]
PERSIANN-CCS	2003 - 2020	0.04	Cloud class system	Satellite	Nguyen et al., 2019 [67]
PERSIANN-CDR	1983 - 2020	0.25	GridSat B1, GPCP	Satellite	Ashouri et al., 2015 [68]
PERSIANN-CCSCDR	2003 - 2020	0.04	GridSat B1, GPCP	Satellite	Sadeghi et al., 2021 [70]
PERSIANN_NOW	2000 - 2020	0.04	IR data, IMERG, WorldClim 2, PERSIANN-CDR, NCEP, GPCP, GSMAP-NOW	Satellite	Nguyen et al., 2020 [69]
PERSIANN	2000 - 2020	0.25	AI on IR and VIS from geostat	Satellite	Nguyen et al., 2019 [67]
TRMM v3B43	1998 - 2019	0.25	PR, TMI, VIRES CERES, LSI	Satellite	Huffman et al., 2007 [71]
UDEL v5.01	1990 - 2017	0.5	GHCN, GSOD	Station	Matsuura and Willmott, 2018 [72]

Valley formation aridifies East Africa and elevates Congo Basin rainfall

https://doi.org/10.1038/s41586-022-05662-5

Received: 9 May 2022

Accepted: 9 December 2022

Published online: 1 March 2023

Check for updates

Callum Munday[1]✉, Nicholas Savage[2], Richard G. Jones[2] & Richard Washington[1]

East African aridification during the past 8 million years is frequently invoked as a driver of large-scale shifts in vegetation[1] and the evolution of new animal lineages, including hominins[2–4]. However, evidence for increasing aridity is debated[5] and, crucially, the mechanisms leading to dry conditions are unclear[6]. Here, numerical model experiments show that valleys punctuating the 6,000-km-long East African Rift System (EARS) are central to the development of dry conditions in East Africa. These valleys, including the Turkana Basin in Kenya, cause East Africa to dry by channelling water vapour towards Central Africa, a process that simultaneously enhances rainfall in the Congo Basin rainforest. Without the valleys, the uplift of the rift system leads to a wetter climate in East Africa and a drier climate in the Congo Basin. Results from climate model experiments demonstrate that the detailed tectonic development of Africa has shaped the rainfall distribution, with profound implications for the evolution of African plant and animal lineages.

East Africa is the driest equatorial landmass on the planet—an unexplained quirk in the classic Hadley model of tropical circulation[7]. Owing to the relatively low annual precipitation (mainly <800 mm year^{-1} in recent decades)[8] and propensity to drought[9], much of East Africa is unable to support dense forest. This is in contrast to the widespread rainforests of central and western equatorial Africa, where annual rainfall is higher, by 150–300% (ref. [10]). Over the past 8 million years, the reduction in East African dense forest, and associated expansion of mixed and grassland habitats[1], has been linked to the evolution of new animal lineages[1], including hominins[2–4]. There is no convincing climatic mechanism explaining this transition and which accounts for continued presence of dense forest in Central Africa. One hypothesis is that drier conditions in East Africa evolved with the uplift of the 6,000-km-long EARS[11,12]; however, mechanisms that relate uplift to rainfall deficits are problematic. In this paper, we demonstrate that the formation of valleys, rather than the uplift alone, is crucial to the low rainfall totals in East Africa. The valleys cause East Africa to dry by channelling water vapour towards Central Africa, a process that simultaneously favours rainfall in the Congo Basin rainforest. Without the valleys, the uplift of the EARS would lead to a wetter climate in East Africa and a drier climate in the Congo Basin.

The EARS separates relatively dry regions of Eastern Africa from wetter Central Africa. It is a leaky barrier to moisture-laden tropical easterly winds, which have been present from at least the Middle Miocene[13]. A series of fault-bounded valleys, oriented east to west, punctuate uplifted topographic domes along the EARS, including the Ethiopian and Kenyan highlands. The valleys are key conduits for water vapour transport from the Indian Ocean to Central Africa, with transport occurring by means of five main easterly low-level jets (LLJs), which accelerate as they are constrained by the valley sides[14,15] (Extended Data Fig. 1). Strong LLJs lead to enhanced water vapour export from East Africa, resulting in low rainfall[15–17]. The valleys—and associated water vapour pathways—probably reached their present-day morphology in the later stages of EARS development (after 10 million years ago (Ma) (ref. [18])). We hypothesize that they are central to the present-day rainfall distribution.

To test this hypothesis, we alter the valleys in a series of 20-year climate model experiments using a 25-km-resolution Pan-African regional climate model from the UK Met Office (see details of the model setup, evaluation and experiments in Methods). The model performs well in its simulation of present-day rainfall and circulation (Extended Data Figs. 1 and 2). Our initial focus is on the Turkana Channel, which is the largest of the fault-bounded valleys. The Turkana Jet—which forms in this valley—is responsible for a large portion of the easterly water vapour transport across the rift system[17]. Phases when the Turkana Jet is strong are associated with drought across the bimodal (two rainy seasons) region of East Africa[15–17] and increased rainfall in downstream regions, including the Ethiopian Highlands and parts of Central Africa[16,19]. We progressively alter the Turkana Channel from a blocked Turkana Channel (A), an incipient channel (B), the control (C; observed topography) and a deeper-than-observed channel (D) (Fig. 1).

The role of the Turkana Channel

Progressive drying occurs in East Africa with the deepening of the Turkana Channel, corresponding to experiments A–D (Fig. 2). Annual mean rainfall is highest in East Africa in the Turkana blocked experiment (Fig. 2a). In this experiment, the model simulates 1,000 mm year^{-1} (+500%) more rainfall in parts of dry northern Kenya and South Sudan compared with the control. Rainfall is lower across rainforest regions of central and western Central Africa, with a reduction of 300 mm year^{-1} near the Atlantic coast. As the channel deepens from experiments B to D, including the control, the upstream region of East Africa becomes progressively drier, whereas the Congo Basin becomes wetter. In the deeper-than-observed experiment (Fig. 2c),

[1]Climate Research Lab, School of Geography and the Environment, University of Oxford, Oxford, UK. [2]Met Office Hadley Centre, Exeter, UK. ✉e-mail: callum.munday@ouce.ox.ac.uk

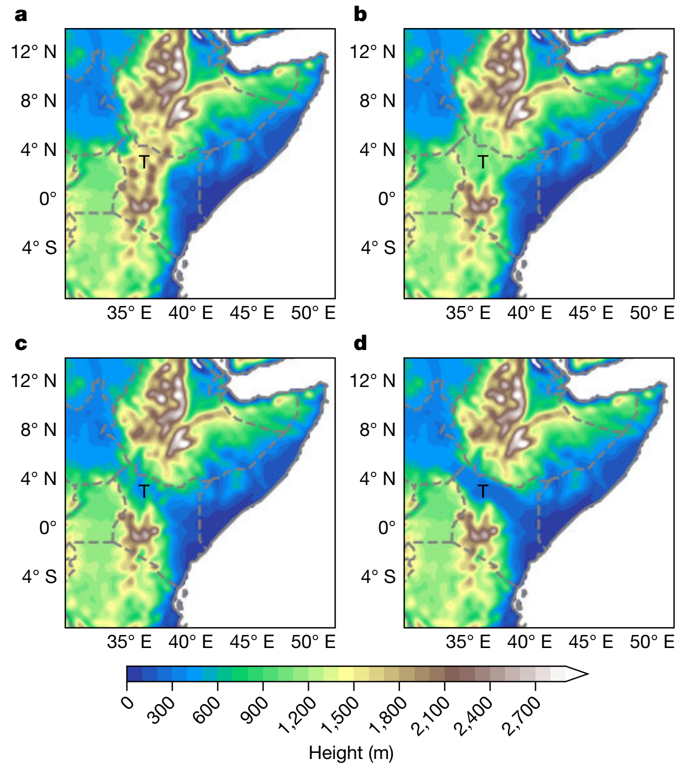


Fig. 1 | Overview of model experiments. Surface altitude (m) for the four Turkana experiments: blocked (**a**), incipient (**b**), control (**c**) and deeper-than-observed channel (**d**). The 'T' marks the location of the Turkana Channel.

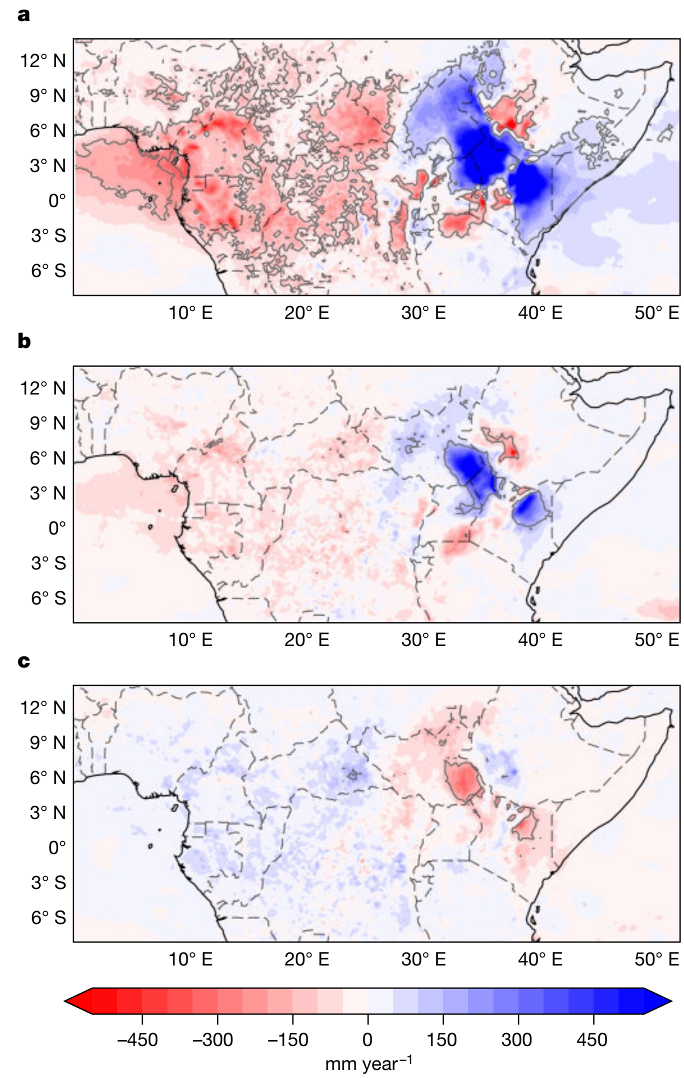


Fig. 2 | The Turkana Channel influences the east–west rainfall gradient in tropical Africa. Annual rainfall anomalies in the experiments compared with control (mm year^{-1}) for Turkana Channel blocked (**a**), incipient (**b**) and deeper-than-observed channel (**c**). Grey contours indicate regions with statistically significant differences between the control and the experiments based on the two-tailed Mann–Whitney *U* test, after controlling for the FDR, following Wilks[31] (see Methods). The P_{FDR} ($n = 20$) values are 0.024, 0.002 and 0.0008 for **a**, **b** and **c**, respectively.

rainfall is reduced in parts of East Africa and South Sudan by 200–400 mm year^{-1} (−30% to 50%). This demonstrates the contribution of the Turkana Channel to the east–west rainfall gradient: the presence of the channel reduces rainfall in East Africa and contributes to the high rainfall totals in Central Africa. The decrease in rainfall over East Africa induced by the Turkana Channel is consistent across all months.

Changes to the integrated water vapour transport (IWVT) help to diagnose the rainfall alterations (Fig. 3). In the blocked experiment (Fig. 3a), there is little easterly water vapour flux from East Africa to Central Africa north of the equator, as the Turkana LLJ is no longer present. IWVT anomalies in the blocked experiment are more than 100 kg m^{-1} s^{-1}. The weakened easterlies are still present in the incipient valley experiment (Fig. 3b). In the deeper-valley experiment (Fig. 3c), there is enhanced easterly water vapour export, associated with the reduced rainfall in East Africa compared with the control. Downstream of the valley in South Sudan, rainfall decreases in the deeper-valley experiment, as there is increased moisture divergence with a stronger jet. In the blocked experiment, rainfall increases in the South Sudan region.

Pan-African effect of valleys

The Turkana Jet is one of five main easterly LLJs that form across the EARS, the others forming in the Limpopo and Zambezi river valleys and across topographic lows in the Tanzanian and Malawi highlands[17,20,21]. These LLJs, and associated valleys, are important for water vapour export, and we posit that they have a similar effect on the east–west rainfall gradient.

To test this hypothesis, we carry out a further experiment (E, no valleys), which blocks all the main topographic lows across the EARS, from the Turkana Channel in the north to the Limpopo River valley in the south (Fig. 4). One can think of this experiment as imposing Andean-like topography across the eastern part of Africa. We find a Pan-African effect on the annual mean rainfall (Fig. 4). Rainfall increases across the entire eastern coastal sector of Africa and in the equatorial Indian Ocean, whereas there is large-scale drying in the continental interior and particularly along the western coast and in South Africa.

In the 'no valleys' experiment, the large-scale wetting is associated with a reduction in water vapour flux through the main valleys (Fig. 4c,d). Meanwhile, in the control experiment, the valleys in the rift system export water vapour from East Africa. The total IWVT across the rift system (see Methods) in the 'no valleys' experiment (2.14×10^8 kg s^{-1}) is 25% less than in the control (2.82×10^8 kg s^{-1}). This leaves more water vapour available for rainfall across Eastern Africa in the 'no valleys' experiment (3.66×10^7 kg s^{-1}) compared with the control simulation (2.73×10^7 kg s^{-1}) (see Methods for moisture budget calculation).

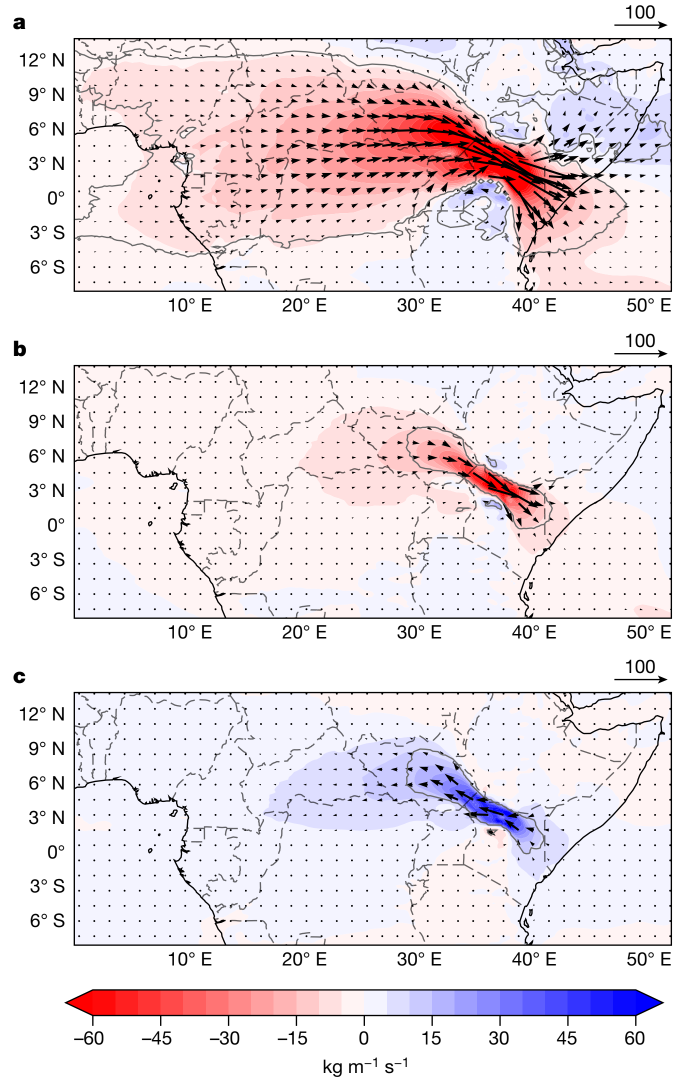


Fig. 3 | Water vapour export from East Africa increases as the Turkana Channel deepens. IWVT anomalies (kg $m^{-1} s^{-1}$; shading) and vectors compared with control (annual mean) for the three Turkana experiments compared with control: Turkana Channel blocked (**a**), incipient (**b**) and deeper-than-observed channel (**c**). Grey contours indicate regions with statistically significant differences (evaluated on the IWVT field) calculated by means of the two-tailed Mann–Whitney *U* test, after controlling for the FDR, following Wilks[31] (see Methods). The P_{FDR} values ($n = 20$) are 0.032, 0.001 and 0.001 for **a**, **b** and **c**, respectively.

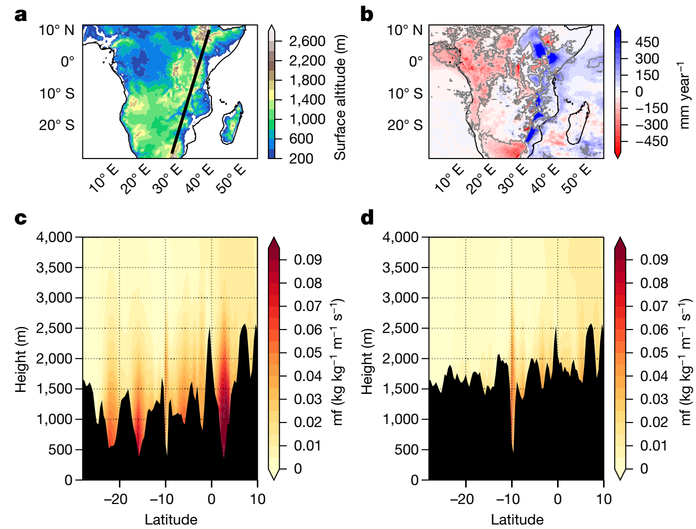


Fig. 4 | Valleys lead to lower rainfall in East Africa, with increased water vapour export across the rift system. a, Surface altitude (m) for the control simulation. **b**, The rainfall anomaly (no valleys − control; mm $year^{-1}$). Vertical cross-sections of the water vapour flux (moisture flux, mf; kg $kg^{-1} m^{-1} s^{-1}$) perpendicular to the black line shown in **a** (positive easterly, negative westerly) for the control (**c**) and 'no valleys' (**d**) experiments. The black filled area indicates the surface height in each experiment. The grey contours in **b** indicate statistically significant differences in rainfall, calculated by means of the two-tailed Mann–Whitney *U* test, after controlling for the FDR, following Wilks[31] (see Methods). The P_{FDR} value ($n = 20$) is 0.022.

Discussion

The presence of valleys within the EARS leads to a drier East Africa and wetter Congo Basin. Previous modelling studies have implicated the rift system in East African dryness[12,22–24], but the shared assumption in these studies is that the uplift alone—which mainly occurred before 10 Ma—causes the aridity. In the 'no valleys' experiment, we show mechanistically that uplift alone does the opposite: the removal of valleys reduces the water vapour export from East Africa, whereas the regions of high topography act as an elevated heat source. This favours convection in East Africa, while necessarily reducing convection in the Congo rainforest, as diagnosed by the reduction in water vapour transport across the rift system. Another important factor in East African aridity, the Somali Jet, strengthens in response to an uninterrupted barrier in the blocked valley experiments (Fig. 3). This is consistent with other modelling studies, which show a weakening of the Somali Jet when the rift system is lowered[22,24].

The expansion of grassland and mixed habitats, at the expense of dense forest, across East Africa in the past 8 million years is linked with broad faunal turnover[1] and is a cornerstone for several hypotheses relating to hominid evolution[2,4,25–28]. The lower rainfall in East Africa owing to the formation of valleys is capable of explaining the large-scale vegetation shift. Although estimates for the timing of valley formation vary between studies and across the EARS, processes occurring in the past 10 Ma have shaped the present-day morphology[11,18]. Other global climate events could have contributed to the expansion of grasslands, including the late Miocene cooling (5–7 Ma)[29] and onset of Northern Hemisphere glaciation (about 2.75 Ma)[30], but the development of valleys is an obvious proximate driver for rainfall change. Today, the presence of valleys is a key explanation for why East Africa is curiously dry for its latitude and prone to drought.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-05662-5.

1. Bobe, R. The evolution of arid ecosystems in eastern Africa. *J. Arid. Environ.* **66**, 564–584 (2006).
2. deMenocal, P. B. African climate change and faunal evolution during the Pliocene–Pleistocene. *Earth Planet. Sci. Lett.* **220**, 3–24 (2004).
3. Maslin, M. A., Shultz, S. & Trauth, M. H. A synthesis of the theories and concepts of early human evolution. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20140064 (2015).
4. Vrba, E. S. in *Evolutionary History of the "Robust" Australopithecines* (ed. Grine, F. E.) 405–426 (Aldine de Gruyter, 1988).
5. Blumenthal, S. A. et al. Aridity and hominin environments. *Proc. Natl Acad. Sci. USA* **114**, 7331–7336 (2017).
6. Levin, N. E. Environment and climate of early human evolution. *Annu. Rev. Earth Planet. Sci.* **43**, 405–429 (2015).
7. Trewartha, G. T. *The Earth's Problem Climates* (Univ. Wisconsin Press, 1961).
8. Yang, W., Seager, R., Cane, M. A. & Lyon, B. The annual cycle of East African precipitation. *J. Clim.* **28**, 2385–2404 (2015).

9. Nicholson, S. E. Climate and climatic variability of rainfall over eastern Africa. *Rev. Geophys.* **55**, 590–635 (2017).
10. Washington, R., James, R., Pearce, H., Pokam, W. M. & Moufouma-Okia, W. Congo Basin rainfall climatology: can we believe the climate models? *Philos. Trans. R. Soc. B Biol. Sci.* **368**, 20120296 (2013).
11. Ring, U. The East African rift system. *Austrian J. Earth Sci.* **107**, 132–146 (2014).
12. Sepulchre, P. et al. Tectonic uplift and Eastern Africa aridification. *Science* **313**, 1419–1423 (2006).
13. Betzler, C. et al. The abrupt onset of the modern South Asian Monsoon winds. *Sci. Rep.* **6**, 29838 (2016).
14. Indeje, M., Semazzi, F. H. M., Xie, L. & Ogallo, L. J. Mechanistic model simulations of the East African climate using NCAR regional climate model: influence of large-scale orography on the Turkana low-level jet. *J. Clim.* **14**, 2710–2724 (2001).
15. Nicholson, S. The Turkana low-level jet: mean climatology and association with regional aridity. *Int. J. Climatol.* **36**, 2598–2614 (2016).
16. Vizy, E. K. & Cook, K. H. Observed relationship between the Turkana low-level jet and boreal summer convection. *Clim. Dyn.* **53**, 4037–4058 (2019).
17. Munday, C., Washington, R. & Har, N. African low-level jets and their importance for water vapor transport and rainfall. *Geophys. Res. Lett.* **48**, e2020GL090999 (2021).
18. Chorowicz, J. The East African rift system. *J. Afr. Earth Sci.* **43**, 379–410 (2005).
19. Viste, E. & Sorteberg, A. Moisture transport into the Ethiopian highlands. *Int. J. Climatol.* **33**, 249–263 (2013).
20. Spavins-Hicks, Z. D., Washington, R. & Munday, C. The Limpopo Low-Level Jet: mean climatology and role in water vapor transport. *J. Geophys. Res. Atmos.* **126**, e2020JD034364 (2021).
21. Barimalala, R., Blamey, R. C., Desbiolles, F. & Reason, C. J. C. The influence of southeastern African river valley jets on regional rainfall. *Clim. Dyn.* **57**, 2905–2920 (2021).
22. Slingo, J., Spencer, H., Hoskins, B., Berrisford, P. & Black, E. The meteorology of the Western Indian Ocean, and the influence of the East African Highlands. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **363**, 25–42 (2005).
23. Prömmel, K., Cubasch, U. & Kaspar, F. A regional climate model study of the impact of tectonic and orbital forcing on African precipitation and vegetation. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **369**, 154–162 (2013).
24. Jung, G., Prange, M. & Schulz, M. Influence of topography on tropical African vegetation coverage. *Clim. Dyn.* **46**, 2535–2549 (2016).
25. DeMenocal, P. B. Plio-Pleistocene African climate. *Science* **270**, 53–59 (1995).
26. Cerling, T. E. et al. Woody cover and hominin environments in the past 6 million years. *Nature* **476**, 51–56 (2011).
27. Maslin, M. A. et al. East African climate pulses and early human evolution. *Quat. Sci. Rev.* **101**, 1–17 (2014).
28. Trauth, M. H. et al. Northern Hemisphere Glaciation, African climate and human evolution. *Quat. Sci. Rev.* **268**, 107095 (2021).
29. Herbert, T. D. et al. Late Miocene global cooling and the rise of modern ecosystems. *Nat. Geosci.* **9**, 843–847 (2016).
30. Hodell, D. A. & Channell, J. E. T. Mode transitions in Northern Hemisphere glaciation: co-evolution of millennial and orbital variability in Quaternary climate. *Clim. Past* **12**, 1805–1828 (2016).
31. Wilks, D. The stippling shows statistically significant grid points. *Bull. Am. Meteorol. Soc.* **97**, 2263–2274 (2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2023

Methods

The model

The model used is the UK Met Office Unified Model set up in the GA7.05 configuration[32]. The model run is from 1990 to 2009 at a 25-km horizontal resolution over a Pan-African domain (34.5° S to 25° N, 13° W to 73° E). The model has a non-hydrostatic semi-implicit and semi-Lagrangian dynamical core[33] and uses a mass-flux convection scheme based on Gregory and Rowntree[34]. Lateral atmospheric boundary conditions are provided by six-hourly fields of temperature, humidity, winds and surface pressure[35] from ERA-Interim reanalysis[36] and daily sea-surface temperatures are prescribed on the basis of the NOAA Daily OISST v2.1 daily reanalysis[37].

To establish the adequacy of the control run, we compare the model data with observed and reanalysis datasets. For precipitation, we compare climatological rainfall with Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) precipitation[38] and Global Precipitation Climatology Project (GPCP) version 2.2 (ref. [39]) (Extended Data Fig. 1). The control run captures the observed distribution of rainfall over Africa, with rainfall hotspots over Madagascar, the Ethiopian Highlands, the Kenyan Highlands, the Rwenzori Mountains and the Cameroonian Highlands. Compared with CHIRPS, the model overestimates rainfall in some of these regions, including in parts of Central Africa and the Ethiopian Highlands. The overestimation of rainfall in the Congo Basin is a common issue across climate models (such as refs. [40,41]). The model performs excellently in East Africa, successfully simulating the low annual rainfall totals.

We compare the model circulation with the latest European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis product (ERA5)[42]. We focus our attention on the zonal and meridional vertically integrated water vapour transport, which is a key diagnostic in Fig. 3. These are calculated on model levels using specific humidity (q) and zonal (u) and meridional (v) winds, as:

$$Qu = \int_{\text{surface}}^{\text{toa}} qu \frac{\mathrm{d}P}{g}$$

$$Qv = \int_{\text{surface}}^{\text{toa}} qv \frac{\mathrm{d}P}{g}$$

$$\text{IWVT} = \sqrt{Qu^2 + Qv^2}$$

in which P is pressure and g is gravitational acceleration (9.81 m s^{-2}), and IWVT is the integrated water vapour transport (kg m^{-1} s^{-1}). The model performs very well in simulating the integrated water vapour flux in comparison with the ERA5 reanalysis (Extended Data Fig. 2). In particular, the model is able to capture the magnitude and distribution of easterly valley jets along the EARS (including the Turkana Jet at about 3° N, 36° E), as well as the southeasterly and southwesterly components of the Somali Jet in the western Indian Ocean.

The experiments

Figure 1 shows the topography in experiments A–D and Fig. 4d shows the topography cross-section in experiment E. For each experiment, we manually edited the valleys in the high-resolution topography file (1-km resolution; NOAA GLOBE digital elevation model[43]) using the open-source GIMP editing software. We altered the high-resolution topography file to address both resolved and sub-grid-scale orographic influences (for example, orographic drag), which are parameterized in the model. The altered high-resolution orography field was used to create new ancillary files (boundary conditions) for each experiment.

The experiments differ from other approaches to understanding orographic effects, which apply a scaling factor to the overall orographic field, for example, by progressively reducing the height of the topography by a given percentage. The approach used here allows us to isolate the effect of the specific geomorphology of interest: the valleys. Because the aim of the experiments is to isolate the effect of valleys on African climate, we do not purport to reproduce specific timespans of Earth's history. The latter is made difficult by a paucity of data on conditions at the time and a lack of agreement between studies on the actual evolution of African topography over the Miocene–Pliocene period[11]. The sensitivity of African climate to valleys, shown in our experiments, suggests that, to clarify temporally linked orographic and climate dynamics, data need to be collected at a sufficient resolution to allow timescales of valley development to be distinguished from broader-scale uplift rates.

Moisture budget calculation

We calculate the moisture budget for the East African region upstream of the rift system shown in Extended Data Fig. 1. On the basis of the Qu and Qv components of the integrated moisture flux, we first use the MetPy cross_section_components function (https://unidata.github.io/MetPy/latest/api/generated/metpy.calc.cross_section_components.html) to find the component of the IWVT normal (inflow) to each segment of the regional boundary (Q_n). We then integrate across each segment separately to find the total segment contribution:

$$Q_{\text{seg}} = \int_{l_0}^{l_n} Q_n \mathrm{d}l$$

in which l is the length of the segment (m) and dl is about 26,000 m (corresponding to the average grid length for the model). We define positive contributions as moisture entering the region and negative contributions as moisture leaving the region. Finally, we sum the contributions from all segments to calculate the moisture budget.

Statistical testing

The two-tailed Mann–Whitney U test is used to evaluate a null hypothesis of no difference between the experiments and the control (n = 20 years). We follow Wilks[31] in applying a constraint on the false discovery rate (FDR) based on the distribution of P-values from each grid point in the composite map. The false discovery control level (α_{FDR}) is set conservatively following Wilks[31] as $2\alpha_{\text{global}}$, in which $\alpha_{\text{global}} = 0.05$, such that $\alpha_{\text{FDR}} = 0.10$. We report the threshold P_{FDR} for each test in the relevant figure captions.

Data availability

Model data arising from this paper used in plotting and the edited high-resolution GLOBE dataset orography files for each experiment are available on publication at https://doi.org/10.5281/zenodo.6956995. ERA5 data were downloaded from https://cds.climate.copernicus.eu/cdsapp#!/home. CHIRPS data are available at https://data.chc.ucsb.edu/products/CHIRPS-2.0/. GPCP data are from https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00979. Data used in base maps for figures are publicly available from https://www.naturalearthdata.com/ and plotted with Cartopy (https://github.com/SciTools/cartopy/archive/v0.11.2.tar.gz). The UK Met Office Unified Model is available for use under license. For further information on how to apply for a license, see http://www.metoffice.gov.uk/research/modelling-systems/unified-model.

Code availability

Code for producing figures is available on publication at https://doi.org/10.5281/zenodo.6956995.

32. Tucker, S. O. et al. Evaluation of a new 12 km regional perturbed parameter ensemble over Europe. *Clim. Dyn.* **58**, 879–903 (2022).
33. Wood, N. et al. An inherently mass-conserving semi-implicit semi-Lagrangian discretization of the deep-atmosphere global non-hydrostatic equations. *Q. J. R. Meteorol. Soc.* **140**, 1505–1520 (2014).
34. Gregory, D. & Rowntree, P. R. A mass flux convection scheme with representation of cloud ensemble characteristics and stability-dependent closure. *Mon. Weather Rev.* **118**, 1483–1506 (1990).
35. Davies, T. Lateral boundary conditions for limited area models. *Q. J. R. Meteorol. Soc.* **140**, 185–196 (2014).
36. Dee, D. P. et al. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **137**, 553–597 (2011).
37. Reynolds, R. W. et al. Daily high-resolution-blended analyses for sea surface temperature. *J. Clim.* **20**, 5473–5496 (2007).
38. Funk, C. et al. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Sci. Data* **2**, 150066 (2015).
39. Adler, R. F. et al. The Global Precipitation Climatology Project (GPCP) monthly analysis (new version 2.3) and a review of 2017 global precipitation. *Atmosphere* **9**, 138 (2018).
40. Creese, A. & Washington, R. Using qflux to constrain modeled Congo Basin rainfall in the CMIP5 ensemble. *J. Geophys. Res. Atmos.* **121**, 13,415–13,442 (2016).
41. Haensler, A., Saeed, F. & Jacob, D. Assessing the robustness of projected precipitation changes over central Africa on the basis of a multitude of global and regional climate projections. *Clim. Change* **121**, 349–363 (2013).
42. Hersbach, H. et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).
43. Hastings, D. A. et al. *The Global Land One-kilometer Base Elevation (GLOBE) Digital Elevation Model, Version 1.0* (National Oceanic and Atmospheric Administration, 1999); http://www.ngdc.noaa.gov/mgg/topo/globe.html.

Acknowledgements We acknowledge the input from and useful discussions with J. Lee-Thorp and R. Bobé (both Oxford University). This article is an output from the REACH programme, financed by UK Aid from the UK Foreign, Commonwealth and Development Office (FCDO) for the benefit of developing countries (programme code 201880). However, the views expressed and information contained in it are not necessarily those of or endorsed by the FCDO, which can accept no responsibility for such views or information or for any reliance placed on them.

Author contributions C.M. designed and ran the model experiments and wrote the manuscript. N.S. helped set up and run the model experiments and wrote part of the Methods section. R.W. contributed to the design of the experiments and edited the manuscript. R.G.J. contributed to the experimental design and edited the manuscript.

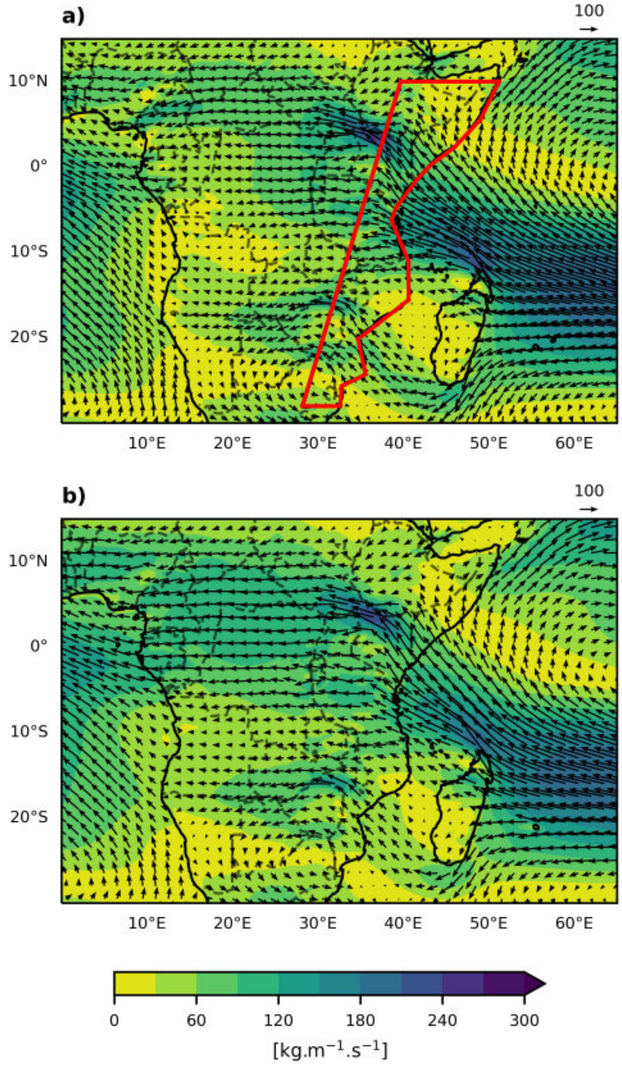
Competing interests The authors declare no competing interests.

Additional information

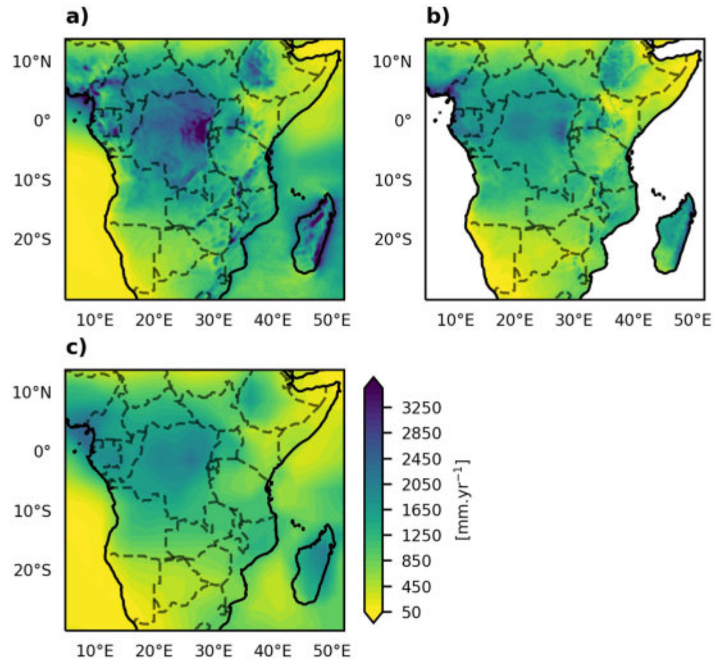
Correspondence and requests for materials should be addressed to Callum Munday.

Peer review information *Nature* thanks Thierry C. Fotso-Nguemo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.



Extended Data Fig. 1 | Water vapour transport in control and reanalysis data. The arrows show the direction and strength at which atmospheric-column integrated water vapour is being transported (kg m^{-1} s^{-1}) in the control experiment (**a**) and ERA5 (**b**). Shading gives the IWVT magnitude (kg m^{-1} s^{-1}) (see Methods). The East African region we use to evaluate the moisture budget is bounded by the red line in **a**. ERA5 data[42] are available at https://cds.climate.copernicus.eu/cdsapp#!/home.



Extended Data Fig. 2 | Evaluation of rainfall in the control simulation. Annual rainfall (mm year^{-1}) in the control (**a**), CHIRPS (**b**) and GPCP version 2.2 (**c**). CHIRPS[38] data are available at https://data.chc.ucsb.edu/products/CHIRPS-2.0/. GPCP data[39] are from https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00979.

Article

Coastal phytoplankton blooms expand and intensify in the 21st century

https://doi.org/10.1038/s41586-023-05760-y

Received: 17 May 2022

Accepted: 25 January 2023

Published online: 1 March 2023

Open access

Check for updates

Yanhui Dai[1,9], Shangbo Yang[1,9], Dan Zhao[1], Chuanmin Hu[2], Wang Xu[3], Donald M. Anderson[4], Yun Li[5], Xiao-Peng Song[6], Daniel G. Boyce[7], Luke Gibson[1], Chunmiao Zheng[1,8] & Lian Feng[1]✉

Phytoplankton blooms in coastal oceans can be beneficial to coastal fisheries production and ecosystem function, but can also cause major environmental problems[1,2]—yet detailed characterizations of bloom incidence and distribution are not available worldwide. Here we map daily marine coastal algal blooms between 2003 and 2020 using global satellite observations at 1-km spatial resolution. We found that algal blooms occurred in 126 out of the 153 coastal countries examined. Globally, the spatial extent (+13.2%) and frequency (+59.2%) of blooms increased significantly ($P < 0.05$) over the study period, whereas blooms weakened in tropical and subtropical areas of the Northern Hemisphere. We documented the relationship between the bloom trends and ocean circulation, and identified the stimulatory effects of recent increases in sea surface temperature. Our compilation of daily mapped coastal phytoplankton blooms provides the basis for global assessments of bloom risks and benefits, and for the formulation or evaluation of management or policy actions.

Phytoplankton blooms are accumulations of microscopic algae in the surface layer of fresh and marine water bodies. Although many blooms can occur naturally, nutrients linked to anthropogenic eutrophication are expected to intensify their frequency globally[2–4]. Many algal blooms are beneficial, fixing carbon at the base of the food chain and supporting fisheries and ecosystems worldwide. However, proliferations of algae that cause harm (termed harmful algal blooms (HABs)) have become a major environmental problem worldwide[5–7]. For instance, the toxins produced by some algal species can accumulate in the food web, causing closures of fisheries as well as illness or mortality of marine species and humans[8–10]. In other cases, the decay of a dense algal bloom can deplete oxygen in bottom waters, forming anoxic 'dead zones' that can cause fish and invertebrate die-offs and ecosystem restructuring, with serious consequences for the well-being of coastal communities[1,11]. Unfortunately, algal bloom frequency and distribution are projected to increase with future climate change[12,13], with some changes causing adverse effects on aquatic ecosystems, fisheries and coastal resources.

Owing to substantial heterogeneity in space and time, algal blooms are challenging to characterize on a large scale[5,14], and thus present knowledge does not allow us to answer one of the most fundamental questions: whether algal blooms have changed in recent decades on a global basis[6,15,16]. For example, although HAB events have been compiled into the UNESCO (United Nations Educational, Scientific, and Cultural Organization) Intergovernmental Oceanographic Commission Harmful Algae Event Database (HAEDAT) globally since 1985, bloom trends are difficult to resolve, owing to inconsistent sampling efforts and the diversity of the eco-environmental or socio-economic effects[6]. Alternatively, satellite data have been used to monitor the ocean surface continuously since 1997 and have enabled bloom detection in many coastal regions[17–19]. However, the technical difficulties in dealing with complex optical features across different types of coastal waters have so far prohibited their application globally[20]. To fill this knowledge gap, we developed a method to map global coastal algal blooms and used this tool to examine satellite images between 2003 and 2020, addressing three fundamental questions: (1) where and how frequently global coastal oceans have been affected by phytoplankton blooms; (2) whether the blooms have expanded or intensified over the past two decades, both globally and regionally; and (3) the identity of the potential drivers.

Mapping global coastal phytoplankton blooms

We generated a satellite-based dataset of phytoplankton bloom occurrence to characterize the spatial and temporal patterns of algal blooms in coastal oceans globally. The dataset was derived using global, 1-km resolution daily observations from the Moderate Resolution Imaging Spectroradiometer (MODIS) onboard NASA's Aqua satellite, and all 0.76 million images acquired by this satellite mission between 2003 and 2020 were used. We developed an automated method to detect phytoplankton blooms using MODIS images (Extended Data Fig. 1) (Methods). In this study, we define a phytoplankton bloom as the accumulation of microscopic algae at the ocean surface that exhibits satellite-detectable fluorescence signals[21]. However, whether a bloom produces toxins or is harmful to humans or the marine environment is not distinguishable from satellite data. We delineated bloom-affected areas (that is, the areas where algal blooms were detected), and enumerated the bloom count

[1]School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen, China. [2]College of Marine Science, University of South Florida, St. Petersburg, FL, USA. [3]Shenzhen Ecological and Environmental Monitoring Center of Guangdong Province, Shenzhen, China. [4]Woods Hole Oceanographic Institution, Woods Hole, MA, USA. [5]School of Marine Science and Policy, College of Earth, Ocean, and Environment, University of Delaware, Lewes, DE, USA. [6]Department of Geographical Sciences, University of Maryland, College Park, MD, USA. [7]Bedford Institute of Oceanography, Fisheries and Oceans Canada, Dartmouth, Nova Scotia, Canada. [8]EIT Institute for Advanced Study, Ningbo, China. [9]These authors contributed equally: Yanhui Dai, Shangbo Yang. ✉e-mail: fengl@sustech.edu.cn

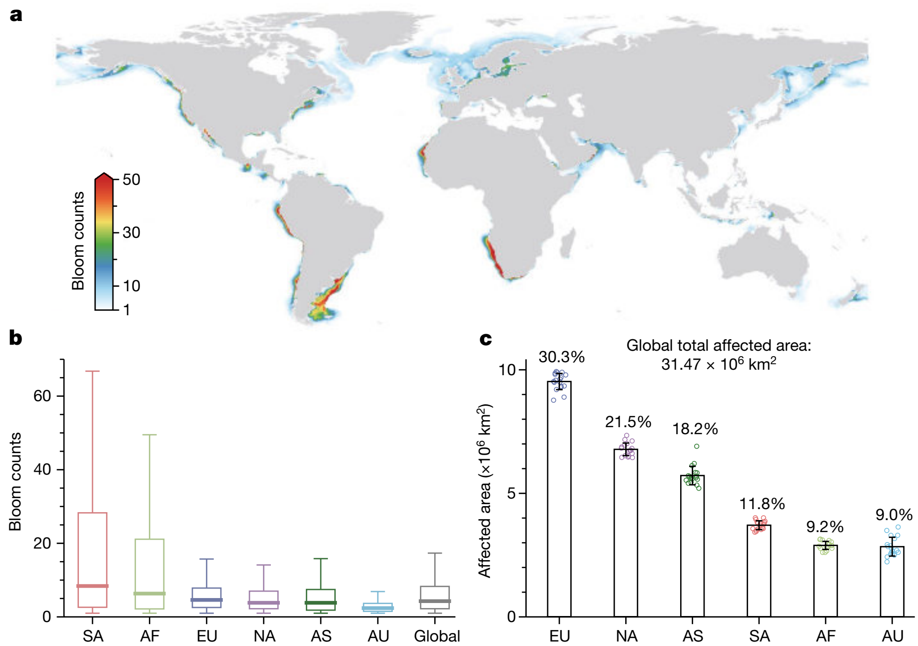


Fig. 1 | Global patterns of coastal phytoplankton blooms between 2003 and 2020. a, The spatial distribution of annual mean bloom count based on daily satellite detections. **b**, Continental and global statistics for annual mean bloom count (South America (SA), n = 3,846,125; Africa (AF), n = 2,516,225; Europe (EU), n = 17,703,949; North America (NA), n = 10,034,286; Asia (AS), n = 5,371,158; Australia (AU), n = 2,781,998 pixel observations). The centre line represents the median value, bottom and top bounds of boxes are first and third quartiles, and the whiskers show a maximum of 1.5 times the interquartile range. **c**, Continental statistics for the long-term annual mean of bloom-affected areas (n = 18 years). The percentages show the corresponding contributions to the global total. The bars represent s.d. Open circles are the affected areas during different years. Map created using Python 3.8.

at the 1-km pixel level (that is, the number of detected blooms per pixel) (Fig. 1). We further estimated the bloom frequency (dimensionless) by integrating the bloom count and affected areas within 1° × 1° grid cells (see Methods), and this metric was used to examine temporal dynamics in bloom intensity. Validation with independent satellite samples selected via several visual inspection techniques showed an overall accuracy level of more than 95% for our method, and comparisons using discrete events in HAEDAT[6] indicated that we successfully identified bloom counts for 79.3% of the historical HAB events in that database (Extended Data Figs. 2–6). We examined phytoplankton blooms in the exclusive economic zones (EEZs) of 153 coastal countries and in 54 large marine ecosystems (LMEs) (Extended Data Fig. 7). Our study area encompasses global continental shelves and outer margins of coastal currents, which offer the majority of marine resources available for human use (see Methods). Out of the 153 coastal countries examined, 126 were observed to have phytoplankton blooms (Fig. 1). The total bloom-affected area was 31.47 million km^2, equivalent to approximately 24.2% of the global land area and 8.6% of the global ocean area, with a median bloom count of 4.3 per year during the past 2 decades (Fig. 1b). Europe (9.52 million km^2—30.3% of the total affected area) and North America (6.78 million km^2—21.5% of the total affected area) contributed the largest bloom areas. By contrast, the most frequent blooms were found around Africa and South America (median bloom counts of more than 6.3 per year). Australia experienced the lowest frequency (2.4 per year) and affected area (2.84 million km^2—9.0% of the total affected area) of blooms.

Phytoplankton blooms occurred frequently in the eastern boundary current systems (that is, California, Benguela, Humboldt and Canary), northeastern USA, Latin America, the Baltic Sea, Northern Black Sea and the Arabian Sea (Fig. 1a). Five LMEs were found with the most frequent blooms (annual median bloom count over 15), including Patagonian Shelf, Northeast US continental Shelf, the Baltic Sea, Gulf of California and Benguela Current (Extended Data Fig. 7). These hotspots are often reported as having a high incidence of algal blooms, some of which are HABs, driven by either coastal upwelling or pronounced anthropogenic nutrient enrichment[9,22–26]. European LMEs showed mostly large proportions of bloom-affected areas, and some also showed frequent bloom occurrences. By contrast, Asian LMEs exhibited mainly infrequent blooms, given their large affected areas. We identified more bloom events in estuarine regions than along coasts in regions without major river discharge ($P < 0.05$; Extended Data Fig. 8), highlighting the critical role of terrestrial nutrient sources in coastal algal blooms[3].

Long-term trends

The total global bloom-affected area has expanded by 3.97 million km^2 (13.2%) between 2003 and 2020, equivalent to 0.14 million km^2 yr^{-1} ($P < 0.05$; Fig. 2). Furthermore, the number of countries with significant bloom expansion was about 1.6 times those with a decreasing trend. The global median bloom frequency showed an increasing rate of 59.2% (+2.19% yr^{-1}, $P < 0.05$) over the observed period. Spatially, areas showing significant increasing trends ($P < 0.05$) in bloom frequency were 77.6% larger than those with the opposite trends (Fig. 2). Globally, our analysis revealed overall consistent fluctuations between the bloom-affected area and bloom frequency between 2003 and 2020 (Fig. 2b). However, there was no significant relationship between bloom extent and frequency in 23 countries and 10 LMEs over the past two decades, underscoring the spatial and temporal variability of algal blooms and the importance of continuous satellite monitoring.

The entire Southern Hemisphere was primarily characterized by increased bloom frequency, although weakened blooms were also sometimes found. In the Northern Hemisphere, the low latitude (<30° N) coasts were mainly featured with strong bloom weakening (Fig. 2a), primarily in the California Current System and the Arabian Sea. Bloom strengthening was found in the northern Gulf of Mexico and the East and South China Seas, albeit at smaller magnitudes. At higher latitudes, weakening blooms were detected mainly in the northeastern North Atlantic and the Okhotsk Sea in the northwestern North Pacific. Globally, the largest increases in bloom frequency were observed in six

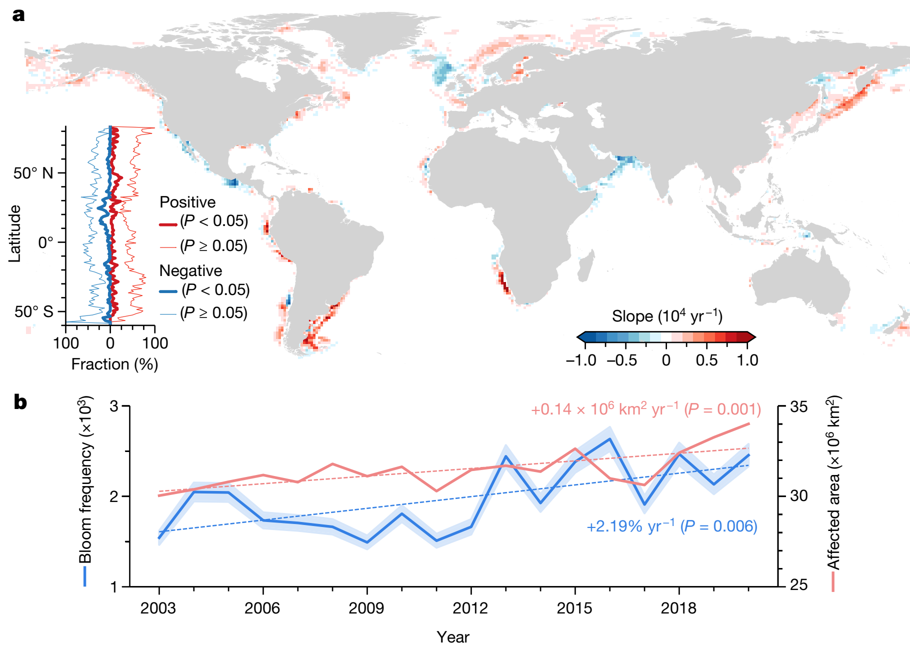


Fig. 2 | Trends of global coastal phytoplankton blooms between 2003 and 2020. a, Spatial patterns of the trends in bloom frequency at a 1° × 1° grid scale. The latitudinal profiles show the fractions of grids with significant and insignificant trends (positive or negative) along the east–west direction. **b**, Interannual variability and trends in annual median bloom frequency and total global bloom-affected area. The linear slopes and *P*-value (two-sided *t*-test) are indicated. The shading associated with the bloom frequency data represents an uncertainty level of 5% in bloom detection. Map created using Python 3.8.

major coastal current systems, including Oyashio (+6.31% yr^{-1}), Alaska (+5.22% yr^{-1}), Canary (+4.28% yr^{-1}), Malvinas (+3.02% yr^{-1}), Gulf Stream (+2.42% yr^{-1}) and Benguela (+2.30% yr^{-1}) (Figs. 2a and 3).

Natural and anthropogenic effects

Increases in sea surface temperature (SST) can stimulate bloom occurrence. We found significant positive correlations ($P < 0.05$) between the annual mean bloom frequency and the coincident SST (SST data were averaged over the growth window of algal blooms within a year (Methods and Extended Data Fig. 9)) in several high-latitude regions (>40° N), such as the Alaska Current ($r = 0.44$), the Oyashio Current ($r = 0.48$) and the Baltic Sea ($r = 0.41$) (Fig. 3). These findings agree with previous studies, in which the bloom-favourable seasons in these temperate seas have been extended under warmer temperatures[27–29]. However, this temperature-based mechanism did not apply to regions with inconsistent trends between SST and bloom frequency, particularly for the substantial bloom weakening in the tropical and subtropical areas (Figs. 2a and 3b).

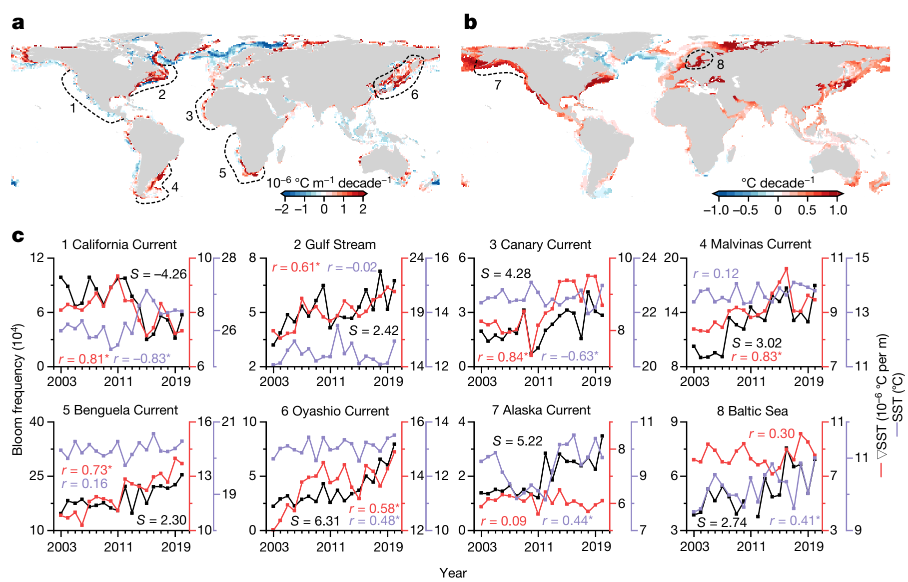


Fig. 3 | Effects of climate change on phytoplankton blooms. a,**b**, Global patterns of trends in SST gradient (**a**) and SST (**b**) from 2003 to 2020. **c**, Long-term changes in bloom frequency in the regions labelled in **a** and **b**, and their relationship to the SST and SST gradient. Linear slope (*S*) of bloom frequency and the correlation coefficient (*r*) between bloom frequency and the SST and the SST gradient (∇SST) are shown. Asterisks indicate statistically significant ($P < 0.05$) correlations. Maps created using ArcMap 10.4 and Python 3.8.

Changes in climate can also affect ocean circulation, altering ocean mixing and the transport of nutrients that drive the growth of marine phytoplankton and bloom formation[30–32]. We used the spatial SST gradient (in °C m^{-1}) as a proxy for the magnitude of oceanic mesoscale currents (the time-varying velocity of kinetic energy (also known as the eddy kinetic energy (EKE))) by following the methods of a previous study[33], and examined its effects on algal blooms (Methods). The trend in the SST gradient appeared more spatially aligned to bloom frequency than SST. We found significant positive correlations ($P < 0.05$) between the SST gradient and bloom frequency for various coastal current systems, including the Canary ($r = 0.84$), Malvinas ($r = 0.83$), California ($r = 0.81$), Benguela ($r = 0.73$), Gulf Stream ($r = 0.61$) and Oyashio ($r = 0.58$) currents.

Trends in bloom frequency in subtropical eastern boundary upwelling regions (the California, Benguela and Canary currents) followed the changes in mesoscale currents (Fig. 3a,c). In the California Current System, the decrease in bloom frequency was probably due to the weakened upwelling (represented by a reduced SST gradient and increased SST) and thus lower nutrient supply[25]. Conversely, the Canary and Benguela currents were characterized by strengthened upwelling and increased bloom frequency. The two western boundary current systems at high latitudes (Malvinas and Oyashio)—although characterized by less pronounced upwelling[34]—exhibited a similar mechanism to the subtropical eastern boundary regions. For the subtropical western boundary Gulf Stream current, the strengthened current jets (manifested as a larger SST gradient) brought more nutrients from the continental shelf[35], triggering more algal blooms. Nevertheless, whether these changes in oceanic mesoscale activities were responses to wind, stratification, the shear of ocean currents or other factors[33] requires region-based investigations.

Global climate events, represented as the multivariate El Niño–Southern Oscillation index[36] (MEI), also showed connections with coastal bloom frequency. The minimum MEI in 2010 (a strong La Niña year) was followed by a low bloom frequency in the following year, and the largest MEI in 2015 (a strong El Niño year) was followed by the strongest bloom frequency in 2016 (Fig. 2b and Extended Data Fig. 10a).

Changes in anthropogenic nutrient enrichment may have also contributed to the trends in blooms[37]. For example, the decline in bloom frequency in the Arabian Sea, without clear links to SST or SST gradient changes, could result from decreased fertilizer use in the surrounding countries (such as Iran) (Extended Data Fig. 10). By contrast, the bloom strengthening in some Asian countries could be attributed to surges in fertilizer use[38]. We examined trends in fertilizer usage (either nitrogen or phosphorus) and bloom frequency and found high positive correlations in China, Iran, Vietnam and the Philippines. Paradoxically, decreased fertilizer uses and increased bloom frequency were identified in some countries, suggesting that nutrient control efforts might have been counterbalanced by the stimulatory effects of climate warming or other factors. Furthermore, the intensified aquaculture industry in Finland, China, Algeria, Guinea, Vietnam, Argentina, Russia and Uruguay may also be linked to their increased bloom incidence, as suggested by the significant positive correlations ($r > 0.5$, $P < 0.05$) between their aquaculture production and bloom frequency. A similar relationship between aquaculture expansion and positive trends in HAB incidence was reported from an analysis of HAEDAT data[6]. However, analogous positive feedbacks for fertilizer or aquaculture were not found in many other countries. Thus, an ecosystem model incorporating terrestrial and oceanic nutrient transport and nutrient–plankton relationships of different species[39] is required to quantify the contributions of natural and anthropogenic factors to algal blooms[14].

Future implications

We acknowledge that our criteria for a detectable bloom event is operationally defined by sensor sensitivities and other factors, and that the bloom count metric used here may underestimate algal bloom incidence, particularly compared to harmful events entered in HAEDAT. For example, in a recent global analysis of the HAEDAT events, Hallegraeff et al.[6] report a dozen or more events per year for each of nine regions over a 33-year study period, compared to the global median bloom count of 4.3 in this study. There are several possible explanations for this discrepancy, such as the many low-cell-concentration HABs that are not detectable from space but that can still cause harm, as well as sensor sensitivities and algorithm thresholds. Furthermore, our bloom count was averaged over all 1-km pixels within the EEZs, whereas HAEDAT entries are based on discrete sampling regions. This underestimation does not, however, alter the trends and other conclusions of this study, as the metrics used here were constant across time and space. Underestimates would have been uniform across regions globally. In this regard, it is of note that the study of Hallegraeff et al.[6] found a significant link between the number of HAEDAT events over time and the global expansion of aquaculture production, similar to findings in our study.

The major contribution of our study is to provide a spatially and temporally consistent characterization of global coastal algal blooms between 2003 and 2020. Globally, increasing trends in algal bloom area and frequency are apparent. Regionally, however, trends were non-uniform owing to the compounded effects of changes in climate (such as changes in SST and SST gradient and climate extremes), anthropogenic eutrophication and aquaculture development. Our daily mapping of bloom events offers valuable baseline information to understand the mechanisms underlying the formation, maintenance, and dissipation of algal blooms[5,40]. This could aid in developing forecasting models (on either global or regional scales) that can help minimize the consequences of harmful blooms, and can also help in policy decisions relating to the control of nutrient discharges and other HAB-stimulatory factors. Noting again that many blooms are beneficial, particularly in terms of their positive effects on ecosystems as well as on wild and farmed fisheries, the results here can also contribute toward policies and management actions that sustain those beneficial blooms.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-05760-y.

1. Breitburg, D. et al. Declining oxygen in the global ocean and coastal waters. *Science* **359**, eaam7240 (2018).
2. Anderson, D. M. Turning back the harmful red tide. *Nature* **388**, 513–514 (1997).
3. Beman, J. M., Arrigo, K. R. & Matson, P. A. Agricultural runoff fuels large phytoplankton blooms in vulnerable areas of the ocean. *Nature* **434**, 211–214 (2005).
4. Heisler, J. et al. Eutrophication and harmful algal blooms: a scientific consensus. *Harmful Algae* **8**, 3–13 (2008).
5. Anderson, D. M., Cembella, A. D. & Hallegraeff, G. M. Progress in understanding harmful algal blooms: paradigm shifts and new technologies for research, monitoring, and management. *Annu. Rev. Mar. Sci.* **4**, 143–176 (2012).
6. Hallegraeff, G. M. et al. Perceived global increase in algal blooms is attributable to intensified monitoring and emerging bloom impacts. *Commun. Earth Environ.* **2**, 117 (2021).
7. Smith, V. H. Eutrophication of freshwater and coastal marine ecosystems a global problem. *Environ. Sci. Pollut. Res.* **10**, 126–139 (2003).
8. Fleming, L. E. et al. Review of Florida red tide and human health effects. *Harmful Algae* **10**, 224–233 (2011).
9. Richlen, M. L., Morton, S. L., Jamali, E. A., Rajan, A. & Anderson, D. M. The catastrophic 2008–2009 red tide in the Arabian Gulf region, with observations on the identification and phylogeny of the fish-killing dinoflagellate *Cochlodinium polykrikoides*. *Harmful Algae* **9**, 163–172 (2010).
10. Hallegraeff, G. & Bolch, C. Unprecedented toxic algal blooms impact on Tasmanian seafood industry. *Microbiol. Aust.* **37**, 143–144 (2016).
11. Diaz, R. J. & Rosenberg, R. Spreading dead zones and consequences for marine ecosystems. *Science* **321**, 926–929 (2008).
12. Barton, A. D., Irwin, A. J., Finkel, Z. V. & Stock, C. A. Anthropogenic climate change drives shift and shuffle in North Atlantic phytoplankton communities. *Proc. Natl Acad. Sci. USA* **113**, 2964–2969 (2016).
13. Gobler, C. J. Climate change and harmful algal blooms: insights and perspective. *Harmful Algae* **91**, 101731 (2020).

14. Zohdi, E. & Abbaspour, M. Harmful algal blooms (red tide): a review of causes, impacts and approaches to monitoring and prediction. *Int. J. Environ. Sci. Technol.* **16**, 1789–1806 (2019).
15. Wells, M. L. et al. Future HAB science: directions and challenges in a changing climate. *Harmful Algae* **91**, 101632 (2020).
16. Rabalais, N. N., Turner, R. E., Díaz, R. J. & Justić, D. Global change and eutrophication of coastal waters. *ICES J. Mar. Sci.* **66**, 1528–1537 (2009).
17. Blondeau-Patissier, D., Gower, J. F., Dekker, A. G., Phinn, S. R. & Brando, V. E. A review of ocean color remote sensing methods and statistical techniques for the detection, mapping and analysis of phytoplankton blooms in coastal and open oceans. *Prog. Oceanogr.* **123**, 123–144 (2014).
18. Wolny, J. L. et al. Current and future remote sensing of harmful algal blooms in the Chesapeake Bay to support the shellfish industry. *Front. Mar. Sci.* **7**, 337 (2020).
19. Stumpf, R. P. et al. Monitoring *Karenia brevis* blooms in the Gulf of Mexico using satellite ocean color imagery and other data. *Harmful Algae* **2**, 147–160 (2003).
20. Bernard, S., Kudela, R. M., Robertson Lain, L. & Pitcher, G. *Observation of Harmful Algal Blooms with Ocean Colour Radiometry* http://dx.doi.org/10.25607/OBP-1042 (IOCCG, 2021).
21. Hu, C. et al. Red tide detection and tracing using MODIS fluorescence data: a regional example in SW Florida coastal waters. *Remote Sens. Environ.* **97**, 311–321 (2005).
22. Andersen, J. H. et al. Long-term temporal and spatial trends in eutrophication status of the Baltic Sea. *Biol. Rev.* **92**, 135–149 (2017).
23. Gómez, F. & Boicenco, L. An annotated checklist of dinoflagellates in the Black Sea. *Hydrobiologia* **517**, 43–59 (2004).
24. Townsend, D. W., Pettigrew, N. R. & Thomas, A. C. Offshore blooms of the red tide dinoflagellate, *Alexandrium* sp., in the Gulf of Maine. *Cont. Shelf Res.* **21**, 347–369 (2001).
25. Pitcher, G. C., Figueiras, F. G., Hickey, B. M. & Moita, M. T. The physical oceanography of upwelling systems and the development of harmful algal blooms. *Prog. Oceanogr.* **85**, 5–32 (2010).
26. López-Cortés, D. J. et al. The state of knowledge of harmful algal blooms of *Margalefidinium polykrikoides* (a.k.a. *Cochlodinium polykrikoides*) in Latin America. *Front. Mar. Sci.* **6**, 463 (2019).
27. Anderson, D. M. et al. Evidence for massive and recurrent toxic blooms of *Alexandrium catenella* in the Alaskan Arctic. *Proc. Natl Acad. Sci. USA* **118**, e2107387118 (2021).
28. Griffith, A. W., Doherty, O. M. & Gobler, C. J. Ocean warming along temperate western boundaries of the Northern Hemisphere promotes an expansion of *Cochlodinium polykrikoides* blooms. *Proc. R. Soc. B* **286**, 20190340 (2019).
29. Conley, D. J. Save the Baltic Sea. *Nature* **486**, 463–464 (2012).
30. Mahadevan, A., D'Asaro, E., Lee, C. & Perry, M. J. Eddy-driven stratification initiates North Atlantic spring phytoplankton blooms. *Science* **337**, 54–58 (2012).
31. Chelton, D. B., Gaube, P., Schlax, M. G., Early, J. J. & Samelson, R. M. The influence of nonlinear mesoscale eddies on near-surface oceanic chlorophyll. *Science* **334**, 328–332 (2011).
32. Boyce, D. G., Petrie, B., Frank, K. T., Worm, B. & Leggett, W. C. Environmental structuring of marine plankton phenology. *Nat. Ecol. Evol.* **1**, 1484–1494 (2017).
33. Martínez-Moreno, J. et al. Global changes in oceanic mesoscale currents over the satellite altimetry record. *Nat. Clim. Change* **11**, 397–403 (2021).
34. Kämpf, J. & Chapman, P. in *Upwelling Systems of the World* 31–65 (Springer, 2016).
35. Lee, T. N., Yoder, J. A. & Atkinson, L. P. Gulf Stream frontal eddy influence on productivity of the southeast US continental shelf. *J. Geophys. Res.* **96**, 22191–22205 (1991).
36. Wolter, K. & Timlin, M. S. El Niño/Southern Oscillation behaviour since 1871 as diagnosed in an extended multivariate ENSO index (MEI. ext). *Int. J. Climatol.* **31**, 1074–1087 (2011).
37. Glibert, P. M. & Burford, M. A. Globally changing nutrient loads and harmful algal blooms: recent advances, new paradigms, and continuing challenges. *Oceanography* **30**, 58–69 (2017).
38. Lu, C. & Tian, H. Global nitrogen and phosphorus fertilizer use for agriculture production in the past half century: shifted hot spots and nutrient imbalance. *Earth Syst. Sci. Data* **9**, 181–192 (2017).
39. Falkowski, P. G., Barber, R. T. & Smetacek, V. Biogeochemical controls and feedbacks on ocean primary production. *Science* **281**, 200–206 (1998).
40. Wells, M. L. et al. Harmful algal blooms and climate change: Learning from the past and present to forecast the future. *Harmful Algae* **49**, 68–93 (2015).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

Methods

Data sources

MODIS on the Aqua satellite provides a global coverage within 1–2 days. All images acquired by this satellite mission from January 2003 to December 2020 were used in our study to detect global coastal phytoplankton blooms, with a total of 0.76 million images. MODIS Level-1A images were downloaded from the Ocean Biology Distributed Active Archive Center (OB.DAAC) at NASA Goddard Space Flight Center (GSFC), and were subsequently processed with SeaDAS software (version 7.5) to obtain Rayleigh-corrected reflectance (R_{rc} (dimensionless), which was converted using the rhos (in sr^{-1}) product (rhos × π) from SeaDAS)[41], remote sensing reflectance (R_{rs} (sr^{-1})) and quality control flags (l2_flags). If a pixel was flagged by any of the following, it was then removed from phytoplankton bloom detection: straylight, cloud, land, high sunglint, high solar zenith angle and high sensor zenith angle (https://oceancolor.gsfc.nasa.gov/atbd/ocl2flags/). MODIS level-3 product for aerosol optical thicknesses (AOT) at 869 nm was also obtained from OB.DAAC NASA GSFC (version R2018.0), which was used to examine the impacts of aerosols on bloom trends.

We examined the algal blooms in the EEZs of 153 ocean-bordering countries (excluding the EEZs in the Caspian Sea or around the Antarctic), 126 of which were found with at least one bloom in the past two decades. The EEZ dataset is available at https://www.marineregions.org/download_file.php?name=World_EEZ_v11_20191118.zip. The EEZs are up to 200 nautical miles (or 370 km) away from coastlines, which include all continental shelf areas and offer the majority of marine resources available for human use. Regional statistics of algal blooms were also performed for LMEs. LMEs encompass global coastal oceans and outer edges of coastal currents areas, which are defined by various distinct features of the oceans, including hydrology, productivity, bathymetry and trophically dependent populations[42]. Of the 66 LMEs identified globally, we excluded the Arctic and Antarctic regions and examined 54 LMEs. The boundaries of LMEs were obtained from https://www.sciencebase.gov/catalog/item/55c7772 2e4b08400b1fd8244.

We used HAEDAT to validate our satellite-detected phytoplankton blooms in terms of presence or absence. The HAEDAT dataset (http://haedat.iode.org) is a collection of records of HAB events, maintained under the UNESCO Intergovernmental Oceanographic Commission and with data archives since 1985. For each HAB event, the HAEDAT records its bloom period (ranging from days to months) and geolocation. We merged duplicate entries when both the recorded locations and times of the HAEDAT events were very similar to one another, and a total number of 2,609 HAEDAT events were ultimately selected between 2003 and 2020.

We used the ¼° resolution National Oceanic and Atmospheric Administration Optimum Interpolated SST (v. 2.1) data to examine the potential simulating effects of warming on the global phytoplankton trends. We also estimated the SST gradients following the method of Martínez-Moreno[33]. As detailed in ref. [33], the SST gradient can be used as a proxy for the magnitude of oceanic mesoscale currents (EKE). We used the SST gradient to explore the effects of ocean circulation dynamics on algal blooms.

Fertilizer uses and aquaculture production for different countries was used to examine the potential effects of nutrient enrichment from humans on global phytoplankton bloom trends. Annual data between 2003 and 2019 on synthetic fertilizer use, including nitrogen and phosphorus, are available from https://ourworldindata.org/fertilizers. Annual aquaculture production includes cultivated fish and crustaceans in marine and inland waters, and sea tanks, and the data between 2003 and 2018 are available from https://ourworldindata.org/grapher/aquaculture-farmed-fish-production.

The MEI, which combines various oceanic and atmospheric variables[36], was used to examine the connections between El Niño–Southern Oscillation activities and marine phytoplankton blooms. The dataset is available from https://psl.noaa.gov/enso/mei/.

Development of an automated bloom detection method

A recent study by the UNESCO Intergovernmental Oceanographic Commission revealed that globally reported HAB events have increased[6]. However, such an overall increasing trend was found to be highly correlated with recently intensified sampling efforts[6]. Once this potential bias was accounted for by examining the ratio between HAB events to the number of samplings[5], there was no significant global trend in HAB incidence, though there were increases in certain regions. With synoptic, frequent, and large-scale observations, satellite remote sensing has been extensively used to monitor algal blooms in oceanic environments[17–19]. For example, chlorophyll *a* (Chl*a*) concentrations, a proxy for phytoplankton biomass, has been provided as a standard product by NASA since the proof-of-concept Coastal Zone Color Scanner (1978–1986) era[43,44]. The current default algorithm used to retrieve Chl*a* products is based on the high absorption of Chl*a* at the blue band[45,46], which often shows high accuracy in the clear open oceans but high uncertainties in coastal waters. This is because, in productive and dynamic coastal oceans, the absorption of Chl*a* in the blue band can be obscured by the presence of suspended sediments and/or coloured dissolved organic matter (CDOM)[47]. To address this problem, various regionalized Chl*a* algorithms have been developed[48]. Unfortunately, the concentrations of the water constituents (CDOM, sediment and Chl*a*) can vary substantially across different coastal oceans. As a result, a universal Chl*a* algorithm that can accurately estimate Chl*a* concentrations in global coastal oceans is not currently available.

Alternatively, many spectral indices have been developed to identify phytoplankton blooms instead of quantifying their bloom biomass, including the normalized fluorescence line height[21] (nFLH), red tide index[49] (RI), algal bloom index[47] (ABI), red–blue difference (RBD)[50], *Karenia brevis* bloom index[50] (KBBI) and red tide detection index[51] (RDI). In practice, the most important task for these index-based algorithms is to determine their optimal thresholds for bloom classification. However, such optimal thresholds can be regional-or image-specific[20], due to the complexity of optical features in coastal waters and/or the contamination of unfavourable observational conditions (such as thick aerosols, thin clouds, and so on), making it difficult to apply spectral-index-based algorithms at a global scale.

To circumvent the difficulty in determining unified thresholds for various spectral indices across global coastal oceans, an approach from a recent study to classify algal blooms in freshwater lakes[52] was adopted and modified here. In that study, the remotely sensed reflectance data in three visible bands (red, green and blue) were converted into two-dimensional colour space created by the Commission Internationale del'éclairage (CIE), in which the position on the CIE chromaticity diagram represented the colour perceived by human eyes (Extended Data Fig. 1a). As the algal blooms in freshwater lakes were manifested as greenish colours, the reflectance of bloom-containing pixels was expected to be distributed in the green gamut of the CIE chromaticity diagram; the stronger the bloom, the closer the distance to the upper border of the diagram (the greener the water).

Here, the colour of phytoplankton blooms in the coastal oceans can be greenish, yellowish, brownish, or even reddish[53], owing to the compositions of bloom species (diatoms or dinoflagellates) and the concentrations of different water constituents. Furthermore, the Chl*a* concentrations of the coastal blooms are typically lower than those in inland waters, thus demanding more accurate classification algorithms. Thus, the algorithm proposed by Hou et al.[52] was modified when using the CIE chromaticity space for bloom detection in marine environments. Specifically, we used the following coordinate conversion formulas to obtain the *xy* coordinate values in the CIE colour space:

$$x = X/(X+Y+Z)$$
$$y = Y/(X+Y+Z)$$
$$X = 2.7689R + 1.7517G + 1.1302B \quad (1)$$
$$Y = 1.0000R + 4.5907G + 0.0601B$$
$$Z = 0.0000R + 0.0565G + 5.5943B$$

where R, G and B represent the R_{rc} at 748 nm, 678 nm (fluorescence band) and 667 nm in the MODIS Aqua data, respectively. By contrast, the R, G and B channels used in Hou et al.[52] were the red, green and blue bands. We used the fluorescence band for the G channel because, for a given region, the 678 nm signal increases monotonically with the Chl*a* concentration for blooms of moderate intensity[21], which is similar to the response of greenness to freshwater algal blooms. Thus, the converted y value in the CIE coordinate system represents the strength of the fluorescence. In practice, for pixels with phytoplankton blooms, the converted colours in the chromaticity diagram will be located within the green, yellow or orange–red gamut (see Extended Data Fig. 1a); the stronger the fluorescence signal is, the closer the distance to the upper border of the CIE diagram (larger y value). By contrast, for bloom-free pixels without a fluorescence signal, their converted xy coordinates will be located in the blue or purple gamut. Therefore, we can determine a lower boundary in the CIE two-dimensional coordinate system to separate bloom and non-bloom pixels, similar to the method proposed by Hou et al.[52].

We selected 53,820 bloom-containing pixels from the MODIS R_{rc} data as training samples to determine the boundary of the CIE colour space. These sample points were selected from nearshore waters worldwide where frequent phytoplankton blooms have been reported (Extended Data Fig. 2); the algal species included various species of dinoflagellates and diatoms[20]. A total of 80 images was used, which were acquired from different seasons and across various bloom magnitudes, to ensure that the samples used could almost exhaustively represent the different bloom conditions in the coastal oceans.

We combined the MODIS FLH_{Rrc} (fluorescence line height based on R_{rc}) and enhanced red–green–blue composite (ERGB) to delineate bloom pixels manually. The FLH_{Rrc} image was calculated as:

$$FLH_{Rrc} = R_{rc678} \times F_{678} - [R_{rc667} \times F_{667} + (R_{rc748} \times F_{748} - R_{rc667} \times F_{667}) \times (678 - 667)/(748 - 667)] \quad (2)$$

where R_{rc667}, R_{rc678} and R_{rc748} are the R_{rc} at 667, 678 and 748 nm, respectively, and F_{667}, F_{678} and F_{748} are the corresponding extraterrestrial solar irradiance. ERGB composite images were generated using R_{rc} of three bands at 555 (R), 488 (G) and 443 nm (B). Although phytoplankton-rich and sediment-rich waters have high FLH_{Rrc} values, they appear as darkish and bright features in the ERGB images (Extended Data Fig. 3), respectively[21]. In fact, visual examination with fluorescence signals and ERGB has been widely accepted as a practical way to delineate coastal algal blooms on a limited number of images[21,54,55]. Note that the FLH_{Rrc} here was slightly different from the NASA standard nFLH product[56], as the latter is generated using R_{rs} (corrected for both Rayleigh and aerosol scattering) instead of R_{rc} (with residual effects of aerosols). However, when using the NASA standard algorithm to further perform aerosol scattering correction over R_{rc}, 20.7% of our selected bloom-containing pixels failed to obtain valid R_{rs} (without retrievals or flagged as low quality), especially for those with strong blooms (see examples in Extended Data Fig. 4). Likewise, we also found various nearshore regions with invalid R_{rs} retrievals. By contrast, R_{rc} had valid data for all selected samples and showed more coverage in nearshore coastal waters. The differences between R_{rs} and R_{rc} were because the assumptions for the standard atmospheric correction algorithm do not hold for bloom pixels or nearshore waters with complex optical properties[57]. In fact, R_{rc} has been used as an alternative to R_{rs} in various applications in complex waters[58,59].

We converted the R_{rc} data of 53,820 selected sample pixels into the xy coordinates in the CIE colour space (Extended Data Fig. 1a). As expected, these samples of bloom-containing pixels were located in the upper half of the chromaticity diagram (the green, yellow and orange–red gamut) (Extended Data Fig. 1a). We determined the lower boundary of these sample points in the chromaticity diagram, which represents the lightest colour and thus the weakest phytoplankton blooms; any point that falls above this boundary represents stronger blooms. The method to determine the boundary was similar to Hou et al.[52]: we first binned the sample points according to the x value in the chromaticity diagram and estimated the 1st percentile (Q_1%) of the corresponding Y for each bin; then, we fit the Q_1% using two-order polynomial regression. Sensitivity analysis with $Q_{0.3}$% (the three-sigma value) resulted in minor changes (<1%) in the resulting bloom areas for single images. Notably, sample points were rarely located near white points ($x = 1/3$ and $y = 1/3$, represent equal reflection from three RGB bands) in the diagram, and we used two polynomial regressions to determine the boundaries for x values greater and less than 1/3, which can be expressed as:

$$y_1 = 4.8093x^2 - 3.0958x + 0.8357 \quad x < \frac{1}{3} \quad (3)$$

$$y_2 = 4.9040x^2 - 3.5759x + 0.9862 \quad x > \frac{1}{3} \quad (4)$$

Based on the above, if a pixel's xy coordinate (converted from R_{rc} spectrum) satisfies the conditions of ($x < 1/3$ AND $y > y_1$) or ($x > 1/3$ AND $y > y_2$), it is classified as a 'bloom' pixel.

Depending on the local region and application purpose, the meaning of 'phytoplankton bloom' may differ. Here, for a global application, the pixelwise bloom classification is based on the relationship (represented using the CIE colour space) between R_{rc} in the 667-, 678- and 754-nm bands derived from visual interpretation of the 80 pairs of FLH_{Rrc} and ERGB imagery. Instead of a simple threshold, we used a lower boundary of the sample points in the chromaticity diagram to define a bloom. In simple words, a pixel is classified as a bloom if its fluorescence signal is detectable (the associated xy coordinate in the CIE colour space located above the lower boundary). Histogram of the nFLH values from the 53,820 training pixels demonstrated the minimum value of ~0.02 mW cm^{-2} µm^{-1} (Extended Data Fig. 1a), which is in line with the lower-bound signal of *K. brevis* blooms on the West Florida shelf[21,47]. Note that, such a minimum nFLH is determined from the global training pixels, and it does not necessarily represent a unified lower bound for phytoplankton blooms across the entire globe, especially considering that fluorescence efficiency may be a large variable across different regions. Different regions may have different lower bounds of nFLH to define a bloom, and such variability is represented by the predefined boundary in the CIE chromaticity diagram in our study. Correspondingly, although the accuracy of Chl*a* retrievals may have large uncertainties in coastal waters, the histogram of the 53,820 training pixels shows a lower bound of ~1 mg m^{-3} (Extended Data Fig. 1a). Similarly to nFLH, such a lower bound may not be applicable to all coastal regions, as different regions may have different lower bounds of Chl*a* for bloom definition.

Although the MODIS cloud (generated by SeaDAS with $R_{rc869} < 0.027$) and associated straylight flags can be used to exclude most clouds, we found that residual errors from thin clouds or cloud shadows could affect the spectral shape and cause misclassification for bloom detections. Thus, we designed two spectral indices to remove such effects:

$$\text{Index1} = nR_{rc488} - nR_{rc443} - (nR_{rc555} - nR_{rc443}) \times 0.5 \quad (5)$$

$$\text{Index2} = nR_{rc555} - nR_{rc469} - (nR_{rc645} - nR_{rc469}) \times 0.5 \quad (6)$$

where Index1 and Index2 were used to remove cloud shadows and clouds, respectively. The nR_{rc443}, nR_{rc488} and nR_{rc555} in index1 are the normalized R_{rc}, obtained by normalizing R_{rc488}. Similar calculations were performed for index2. The purpose of normalizations is to eliminate the effect of the absolute magnitude of the reflectance, so that the thresholds of these two indices are influenced by only the relative magnitude (spectral shape). We determined thresholds for Index1 (>0.12) and Index2 (<0.012) through trial-and-error and ensured that the misclassifications caused by residual errors from clouds and cloud shadows could be effectively removed. After applying the cloud/cloud shadow and various other masks that are associated with l2_flags, we obtained an annual mean valid pixel observation (N_{vobs}) of ~2.0×10^5 for global 1° × 1° grid cells, and the fluctuation patterns and trends of N_{vobs}, either annually or seasonally, are different from that of the global bloom frequency and affected areas (see Supplementary Fig. 1).

Assessments of the algorithm performance

In addition to phytoplankton blooms, macroalgal blooms (*Sargassum* and *Ulva*) frequently occur in many coastal oceans[60–63]. To verify whether our CIE-fluorescence algorithm could eliminate such impacts, we compared the spectra between micro-and macroalgal blooms (see Extended Data Fig. 1b). We found that the spectral shapes of *Sargassum* and *Ulva* are substantially different from those of microalgae, particularly for the three bands used for CIE coordinate conversion. The converted *xy* coordinates for macroalgae were located in the purple–red gamut of the CIE diagram, which was far below the predefined boundary (Extended Data Fig. 1). Moreover, our algorithm is not affected by highly turbid waters for the following two reasons: first, extremely high turbidity tends to saturate the MODIS ocean bands[64], which can be easily excluded; second, without a fluorescence peak, the reflectance of unsaturated turbid waters, after conversion to CIE coordinates, will be located below the predefined boundary (see example in Extended Data Fig. 1b). We also confirmed that the spectral shapes of coccolithophore blooms are different from dinoflagellates and diatoms (see example in Extended Data Fig. 1b), and thus they are excluded from our algorithm.

Three different types of validation methods were adopted to demonstrate the reliability of the proposed CIE-fluorescence algorithm for phytoplankton bloom detection in global coastal oceans, including visual inspections of the RGB, ERGB and FLH_{Rrc} images, verifications using independent manually delineated algal blooms, and comparisons with the reported HAB events from the HAEDAT dataset.

First, we selected MODIS Aqua images from different locations where coastal phytoplankton blooms have been recorded in the published literature. We visually compared the RGB, ERGB, and FLH_{Rrc} images, and our algorithm detected bloom patterns (see examples in Extended Data Fig. 3). Comparisons from various images worldwide showed that our algorithm could successfully identify regions with high FLH_{Rrc} values and brownish-to-darkish features on the ERGB images, which can be considered phytoplankton blooms.

Second, we delineated additional 15,466 bloom-containing pixels from 35 images covering different coastal areas, using the same visual inspection and manual delineation method as for the training sample pixels. Moreover, we also selected 14,149 bloom-free pixels (bright or turquoise green colours on ERGB images or low FLH_{Rrc} values) within the same images as bloom-containing images. We applied our algorithm to all these pixels, and compared the algorithm-identified and manually delineated results. Our CIE-fluorescence algorithm showed high values in both producer and user accuracies (92.04% and 98.63%) (Supplementary Table 1), and appeared effective at identifying bloom pixels and excluding false negatives (blooms classified as non-blooms) and false positives (non-blooms classified as blooms).

Third, we validated the satellite-detected phytoplankton blooms using in situ reported HAB events from the HAEDAT dataset. For each HAB event in the HAEDAT dataset, we obtained all MODIS images over the reported bloom period (from days to months). Within each year, we estimated the ratio between the number of satellite images with 'bloom detected' against the number of valid images (see definition above) during the bloom periods across the entire globe (Supplementary Table 1). Moreover, we calculated the number of events with at least one successful satellite bloom detection (N_s), and then estimated the ratio between N_s and the total HAB events for each year. Results showed that substantial amounts (averaged at 51.2%) of satellite observations acquired during the HAB event periods were found with phytoplankton blooms by our algorithm. Overall, 79.3% of the global HAB events were successfully identified by satellite. The discrepancies between satellite and in situ observations could be explained by the following reasons: first, our study focused only on the phytoplankton blooms that are resolvable by satellite fluorescence signals; other types of HAB events in the HAEDAT dataset may not have been detectable by satellite observations, such as events with lower phytoplankton biomass but high toxicity, occurrences at the subsurface layers, or fluorescence signals overwhelmed by suspended sediments[65–67]. Second, although the HAEDAT recorded HAB events could be sustained for long periods, high biomass of surface algae may not have occurred every day within this period due to the changes in stratification, precipitation, wind, vertical migration of cells, and many other factors[68]. Third, the spatial scale of certain HAB events may have been too small to be identified using the 1-km resolution MODIS observations. Fourth, a reduced MODIS satellite observation frequency by the contaminations of clouds and land adjacency effects[69]. Therefore, we believe the underestimations of satellite-detected blooms compared to the in situ reported HAB events were mainly due to inconsistencies between the two observations rather than uncertainties in our algorithm.

Because R_{rc} depends not only on water colour but also on aerosols (type and concentration) and solar and viewing geometry, a sensitivity analysis was used to determine whether such variables could impact bloom detection. Aerosol reflectance (ρ_a) with different AOTs at 869 nm was simulated using the NASA-recommended maritime aerosol model (r75f02, with a relative humidity of 75% and a fine-mode fraction of 2%). Then, ρ_a of each MODIS band was added to R_{rc} images, and the resulting bloom areas with and without added ρ_a were compared. Results showed that even with a change of 0.02 in AOT at 869 nm, the bloom areas showed minor changes (<2%) in the tested images; minor changes were also found when we used different aerosol models to conduct ρ_a simulations[70]. Note that 0.02 represents the high end of the AOT intra-annual variability in coastal oceans (see Extended Data Fig. 5), and the associated interannual changes are much smaller. Thus, the use of R_{rc} instead of the fully atmospherically corrected reflectance R_{rs} could have limited impacts on our detected global bloom trend.

We also tried various index-based algorithms developed in previous studies. However, results showed that all these methods require image-specific thresholds to accurately determine algal bloom boundaries for different coastal regions (see Extended Data Fig. 6). By contrast, although our CIE-fluorescence algorithm may lead to different bloom thresholds for different regions, it can identify bloom pixels without adjusting the coefficients and, therefore, is more suitable for global-scale bloom assessment efforts.

We acknowledge that our satellite-detected algal blooms represent only high amounts of phytoplankton biomass on the ocean surfaces without distinguishing whether such blooms produce toxins or are harmful to marine environments. Furthermore, with only limited spectral information from MODIS, it is difficult to discriminate the phytoplankton species of algal blooms; such information could help to improve our understanding of the impacts of these phytoplankton blooms. However, we expect these challenges to be addressed soon with the scheduled launch of the Plankton, Aerosol, Cloud, ocean Ecosystem (PACE) mission by NASA in 2024, where the hyperspectral measurements over a broad spectrum (350–885 nm) will make species-level classifications possible[71].

Exploring the patterns and trends of global coastal phytoplankton blooms

We applied the CIE-fluorescence algorithm to all MODIS Aqua level-2 R_{rc} images, and a total number of 0.76 million images between 2003 and 2020 were processed. We mapped all detected blooms into 1-km daily scale level-3 composites. The number of bloom counts within a year for each location can be easily enumerated, and the long-term annual mean values were then estimated (Fig. 1a). We further calculated the total global bloom-affected area (the areas where algal blooms were detected at least once) for each year and examined their changes over time (Fig. 2b).

We defined bloom frequency (dimensionless) to represent the density of phytoplankton blooms for a year by integrating the bloom count and bloom-affected areas within 1°×1° grid cells within that year, which is expressed as:

$$\text{Bloom frequency} = \frac{n}{N}\sum_{i=1}^{n} M_i \quad (7)$$

where M_i is the enumerated bloom count for each 1-km resolution pixel in a year within one 1° × 1° grid cell, and n represents the associated number of bloom-affected pixels in the same cell (the number of pixels with $M_i > 0$), and N is the total number of 1-km MODIS pixels in this grid cell. We estimated the bloom frequency for each year between 2003 and 2020, and determined the long-term trend over global EEZs through a linear least-squares regression (see Fig. 2a).

Continental and country-level statistics were performed for bloom count, bloom-affected areas, and bloom frequency (Fig. 1b,c and Supplementary Table 2), using boundaries for the EEZs of different ocean-bordering countries (see above). Similar statistics were also conducted for 54 LMEs (Extended Data Fig. 7 and Supplementary Table 3).

Correlations with SST and SST gradient

To assess the impacts of climate change on long-term trends in coastal phytoplankton blooms, we correlated the annual mean bloom frequency and the associated SST and SST gradient in various coastal current systems for grid cells with significant changes in bloom frequency (Fig. 3c). The SST and SST gradient were averaged over the growth window within a year, assuming that the changes within the growth window, either in water temperatures or ocean circulations, play more important roles in the bloom trends compared to other seasons[32].

We determined the growth window of phytoplankton blooms for each 1° × 1° grid cell (Extended Data Fig. 9a) using the following method: first, we estimated the proportion of cumulative bloom-affected pixels within the grid cells for a year. Second, a generalized additive model[72] was used to determine the shape of the phenological curves (Extended Data Fig. 9b), where a log link function and a cubic cyclic regression spline smoother were applied[73,74]. Third, the timing of maximum bloom-affected areas (TMBAA) was then determined by identifying the inflection point on the bloom growth curve (Extended Data Fig. 9c). To facilitate comparisons across Northern and Southern Hemispheres, the year in the Southern Hemisphere was shifted forward by 183 days (Extended Data Fig. 9c). We characterized the similarity of the bloom growth curve between different grid cells and grouped them into three distinct clusters using a fuzzy c-means cluster analysis method[75,76]. We found uniform distributions of the clusters over large geographic areas. Cluster I is mainly distributed in mid-low latitudes (<45° N and <30° S), where the maximum bloom-affected areas were expected in the early period of the year. Cluster II was mostly found in higher latitudes, with bloom developments (quasi-) synchronized with increases in SST. Cluster III was detected along the coastlines, where the bloom-affected areas increase throughout the entire year. In practice, the growth window for clusters I and III was set as the entire year, and that for cluster II was set from day 150 to day 270 within the year. We further found that the TMBAA for cluster II showed small changes over the entire period (Extended Data Fig. 9d), indicating relatively stable phenological cycles for those phytoplankton blooms[32,77].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The satellite-based dataset of global coastal algal bloom at 1-km resolution and the associated code are available at https://doi.org/10.5281/zenodo.7359262. Source data are provided with this paper.

41. Hu, C. A novel ocean color index to detect floating algae in the global oceans. *Remote Sens. Environ.* **113**, 2118–2129 (2009).
42. Sherman, K. Adaptive management institutions at the regional level: the case of large marine ecosystems. *Ocean Coast. Manag.* **90**, 38–49 (2014).
43. Gordon, H. R., Clark, D. K., Mueller, J. L. & Hovis, W. A. Phytoplankton pigments from the Nimbus-7 coastal zone color scanner: comparisons with surface measurements. *Science* **210**, 63–66 (1980).
44. Moore, J. K. & Abbott, M. R. Phytoplankton chlorophyll distributions and primary production in the Southern Ocean. *J. Geophys. Res.* **105**, 28709–28722 (2000).
45. Hu, C. et al. Improving satellite global chlorophyll *a* data products through algorithm refinement and data recovery. *J. Geophys. Res.* **124**, 1524–1543 (2019).
46. Hu, C., Lee, Z. & Franz, B. Chlorophyll *a* algorithms for oligotrophic oceans: a novel approach based on three-band reflectance difference. *J. Geophys. Res.* **117**, C01011 (2012).
47. Hu, C. & Feng, L. Modified MODIS fluorescence line height data product to improve image interpretation for red tide monitoring in the eastern Gulf of Mexico. *J. Appl. Remote Sens.* **11**, 012003 (2016).
48. Siswanto, E. et al. Empirical ocean-color algorithms to retrieve chlorophyll-*a*, total suspended matter, and colored dissolved organic matter absorption coefficient in the Yellow and East China Seas. *J. Oceanogr.* **67**, 627–650 (2011).
49. Ahn, Y.-H. & Shanmugam, P. Detecting the red tide algal blooms from satellite ocean color observations in optically complex Northeast-Asia coastal waters. *Remote Sens. Environ.* **103**, 419–437 (2006).
50. Amin, R. et al. Novel optical techniques for detecting and classifying toxic dinoflagellate *Karenia brevis* blooms using satellite imagery. *Opt. Express* **17**, 9126–9144 (2009).
51. Shen, F., Tang, R., Sun, X. & Liu, D. Simple methods for satellite identification of algal blooms and species using 10-year time series data from the East China Sea. *Remote Sens. Environ.* **235**, 111484 (2019).
52. Hou, X. et al. Global mapping reveals increase in lacustrine algal blooms over the past decade. *Nat. Geosci.* **15**, 130–134 (2022).
53. Dierssen, H. M., Kudela, R. M., Ryan, J. P. & Zimmerman, R. C. Red and black tides: quantitative analysis of water-leaving radiance and perceived color for phytoplankton, colored dissolved organic matter, and suspended sediments. *Limnol. Oceanogr.* **51**, 2646–2659 (2006).
54. Zhao, J., Temimi, M. & Ghedira, H. Characterization of harmful algal blooms (HABs) in the Arabian Gulf and the Sea of Oman using MERIS fluorescence data. *ISPRS J. Photogramm. Remote Sens.* **101**, 125–136 (2015).
55. Qi, L. et al. Noctiluca blooms in the East China Sea bounded by ocean fronts. *Harmful Algae* **112**, 102172 (2022).
56. Behrenfeld, M. J. et al. Satellite-detected fluorescence reveals global physiology of ocean phytoplankton. *Biogeosciences* **6**, 779–794 (2009).
57. Gordon, H. R. Atmospheric correction of ocean color imagery in the Earth Observing System era. *J. Geophys. Res.* **102**, 17081–17106 (1997).
58. Feng, L., Hou, X., Li, J. & Zheng, Y. Exploring the potential of Rayleigh-corrected reflectance in coastal and inland water applications: a simple aerosol correction method and its merits. *ISPRS J. Photogramm. Remote Sens.* **146**, 52–64 (2018).
59. Feng, L., Hu, C., Han, X., Chen, X. & Qi, L. Long-term distribution patterns of chlorophyll-*a* concentration in China's largest freshwater lake: MERIS full-resolution observations with a practical approach. *Remote Sens.* **7**, 275–299 (2015).
60. Xiao, J. et al. An anomalous bi-macroalgal bloom caused by *Ulva* and *Sargassum* seaweeds during spring to summer of 2017 in the western Yellow Sea, China. *Harmful Algae* **93**, 101760 (2020).
61. Teichberg, M. et al. Eutrophication and macroalgal blooms in temperate and tropical coastal waters: nutrient enrichment experiments with *Ulva* spp. *Glob. Change Biol.* **16**, 2624–2637 (2010).
62. Viaroli, P. et al. Nutrient and iron limitation to Ulva blooms in a eutrophic coastal lagoon (Sacca di Goro, Italy). *Hydrobiologia* **550**, 57–71 (2005).
63. Dierssen, H. M., Chlus, A. & Russell, B. Hyperspectral discrimination of floating mats of seagrass wrack and the macroalgae *Sargassum* in coastal waters of Greater Florida Bay using airborne remote sensing. *Remote Sens. Environ.* **167**, 247–258 (2015).
64. Hu, C. et al. Dynamic range and sensitivity requirements of satellite ocean color sensors: learning from the past. *Appl. Opt.* **51**, 6045–6062 (2012).
65. Trainer, V. L. et al. Pelagic harmful algal blooms and climate change: lessons from nature's experiments with extremes. *Harmful Algae* **91**, 101591 (2020).
66. Mardones, J. I. et al. Disentangling the environmental processes responsible for the world's largest farmed fish-killing harmful algal bloom: Chile, 2016. *Sci. Total Environ.* **766**, 144383 (2021).

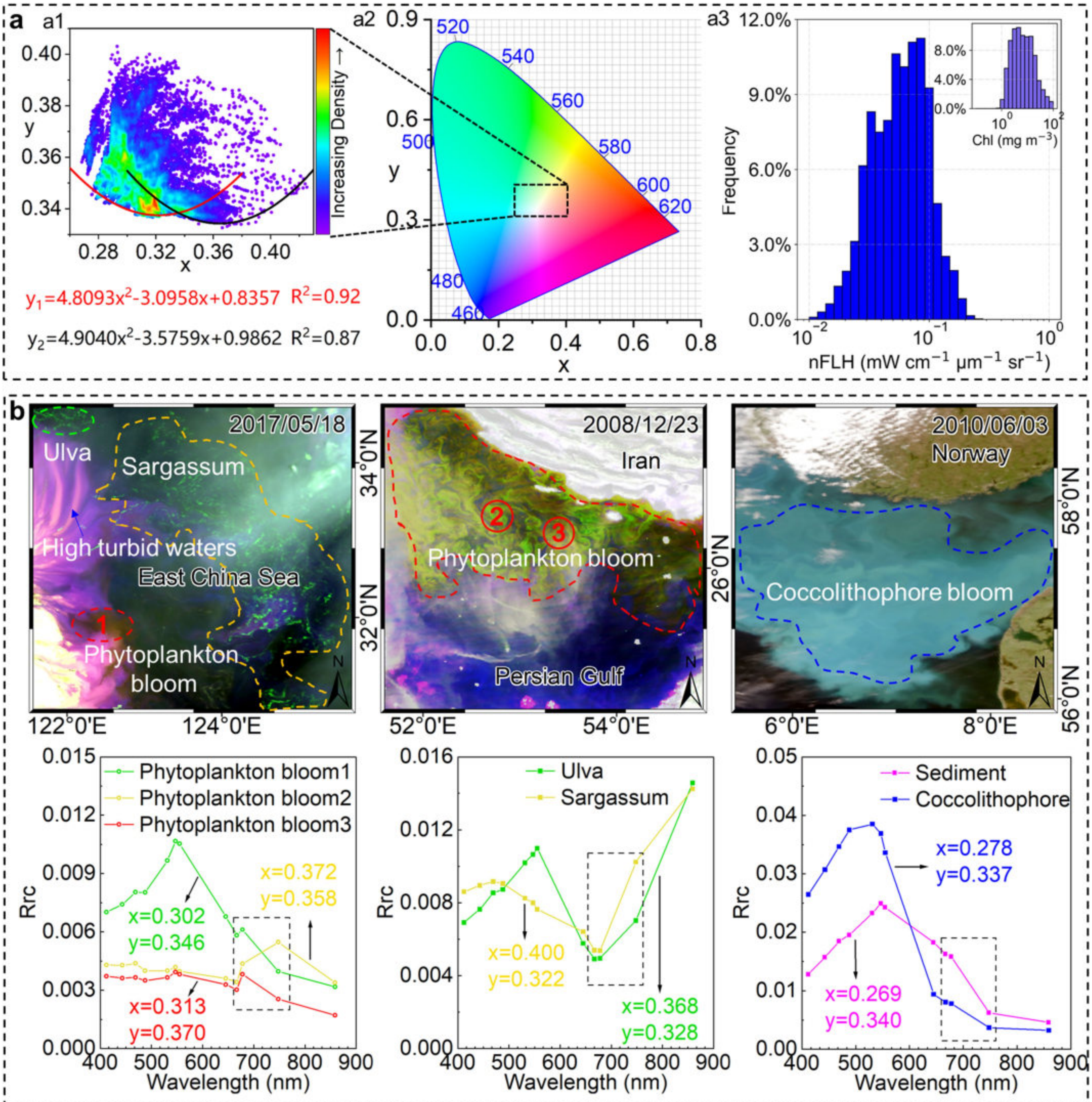
67. Gilerson, A. et al. Fluorescence component in the reflectance spectra from coastal waters. II. Performance of retrieval algorithms. *Opt. Express* **16**, 2446–2460 (2008).
68. Lee, J. H., Harrison, P. J., Kuang, C. & Yin, K. in *The Environment in Asia Pacific Harbours* (ed. Wolanski, E.) 187–206 (Springer, 2006).
69. Feng, L. & Hu, C. Cloud adjacency effects on top-of-atmosphere radiance and ocean color data products: a statistical assessment. *Remote Sens. Environ.* **174**, 301–313 (2016).
70. Ahmad, Z. et al. New aerosol models for the retrieval of aerosol optical thickness and normalized water-leaving radiances from the SeaWiFS and MODIS sensors over coastal regions and open oceans. *Appl. Opt.* **49**, 5545–5560 (2010).
71. Werdell, P. J. et al. The Plankton, Aerosol, Cloud, Ocean Ecosystem mission: status, science, advances. *Bull. Am. Meteorol. Soc.* **100**, 1775–1794 (2019).
72. Hastie, T. J. *Generalized Additive Models* (Routledge, 2017).
73. Wood, S. N. *Generalized Additive Models: An Introduction with R* (CRC press, 2017).
74. Macgregor, C. J. et al. Climate-induced phenology shifts linked to range expansions in species with multiple reproductive cycles per year. *Nat. Commun.* **10**, 4455 (2019).
75. Bezdek, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms* (Springer, 2013).
76. Bi, S. et al. Optical classification of inland waters based on an improved fuzzy C-means method. *Opt. Express* **27**, 34838–34856 (2019).
77. Kheireddine, M., Mayot, N., Ouhssain, M. & Jones, B. H. Regionalization of the Red Sea based on phytoplankton phenology: a satellite analysis. *J. Geophys. Res.* **126**, e2021JC017486 (2021).

Acknowledgements We thank NASA for providing global MODIS satellite images, and the Intergovernmental Oceanographic Commission (IOC) of UNESCO for providing the HAEDAT dataset. L.F. was supported by the National Natural Science Foundation of China (no. 41890852, 42271322 and 41971304). D.M.A. was supported by the Woods Hole Center for Oceans and Human Health (National Science Foundation grant OCE-1840381 and National Institutes of Health grants NIEHS-1P01-ES028938-01).

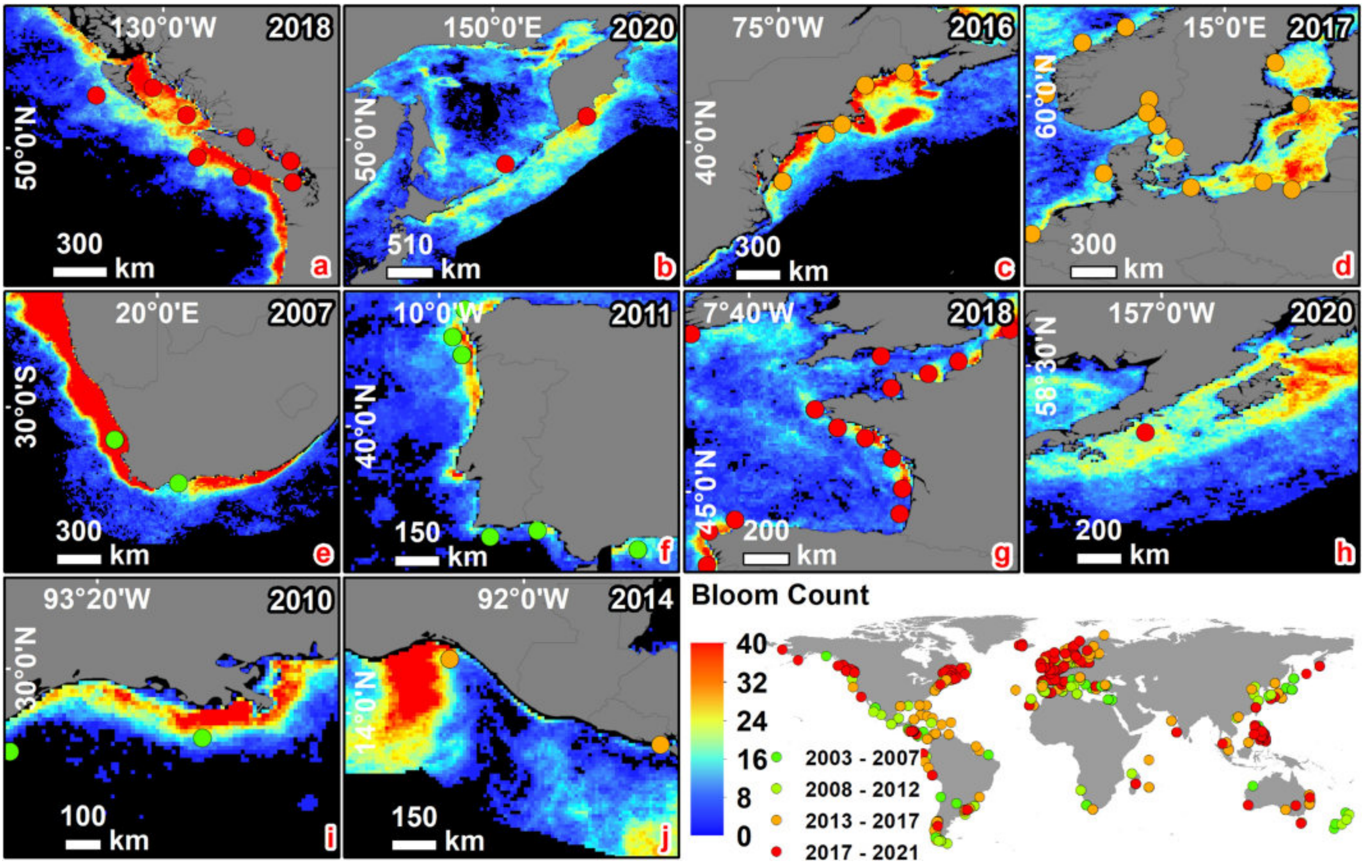
Author contributions Y.D. and S.Y.: methodology, data processing and analyses, and writing. L.F.: conceptualization, methodology, funding acquisition, supervision and writing. D.Z.: data processing and analysis. C.H., W.X., D.M.A., Y.L., X.-P.S., D.G.B., L.G. and C.Z. participated in interpreting the results and refining the manuscript.

Competing interests The authors declare no competing interests.

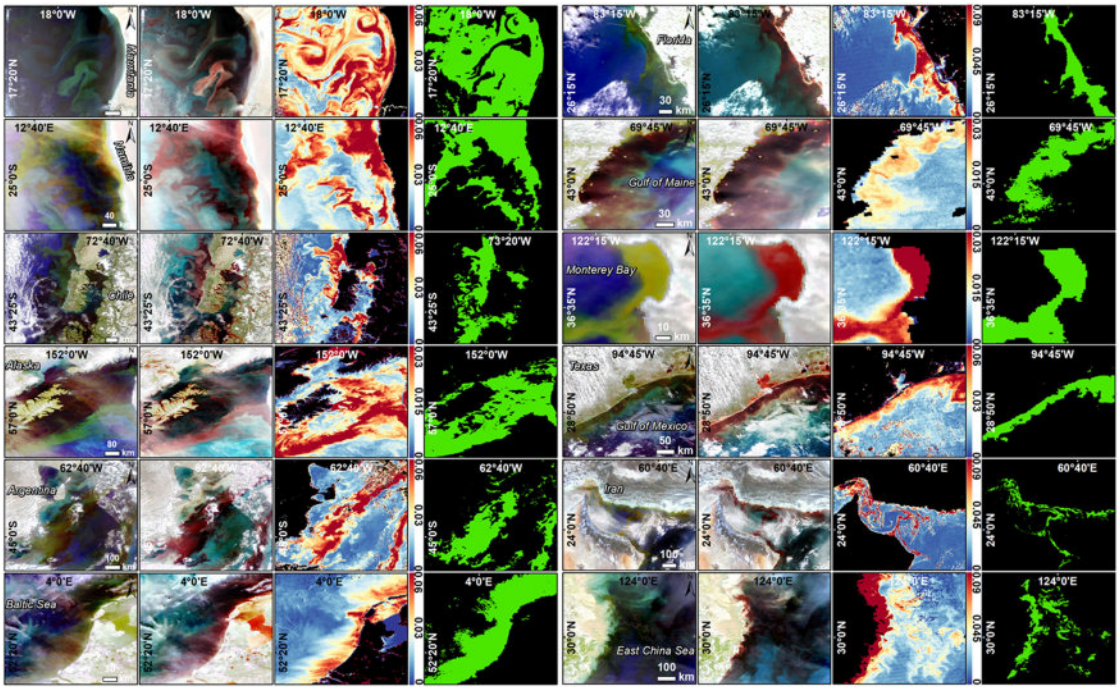
Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-05760-y.
Correspondence and requests for materials should be addressed to Lian Feng.
Peer review information *Nature* thanks Bryan Franz, Bingkun Luo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.
Reprints and permissions information is available at http://www.nature.com/reprints.



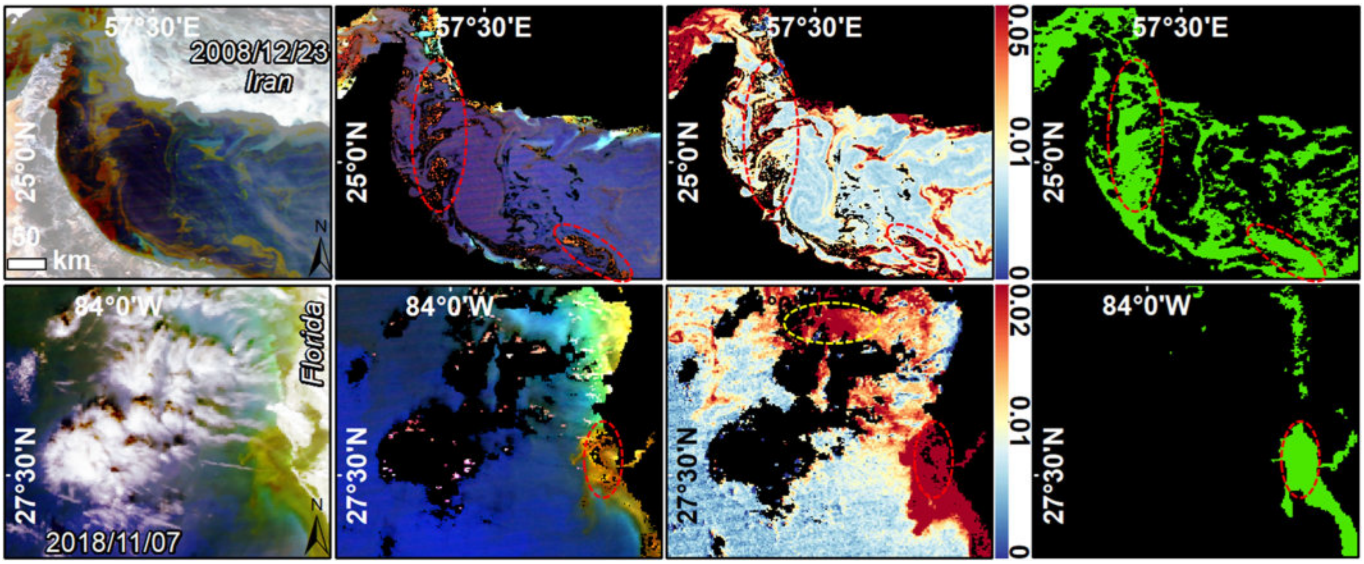
Extended Data Fig. 1 | Development of the CIE-fluorescence algorithm to detect phytoplankton blooms using MODIS satellite imagery. (**a**). A1: The density plot of manually delineated bloom-containing pixels in the CIE coordinate system (*n* = 53,820), and their distribution in the CIE color space (box in A2). A3: Histograms of nFLH and Chla for the delineated pixels, obtained using NASA standard algorithms[47,57]. (**b**) MODIS true color composites and selected spectra for phytoplankton blooms, macroalgal blooms (*Ulva* and *Sargassum*), coccolithophore blooms, and sediment-rich turbid waters. The x-y numbers indicate their corresponding positions in the CIE coordinate system. The black rectangular boxes in the three lower panels highlight different spectral shapes between phytoplankton blooms and other features near the fluorescence band. Maps created using ArcMap 10.4.



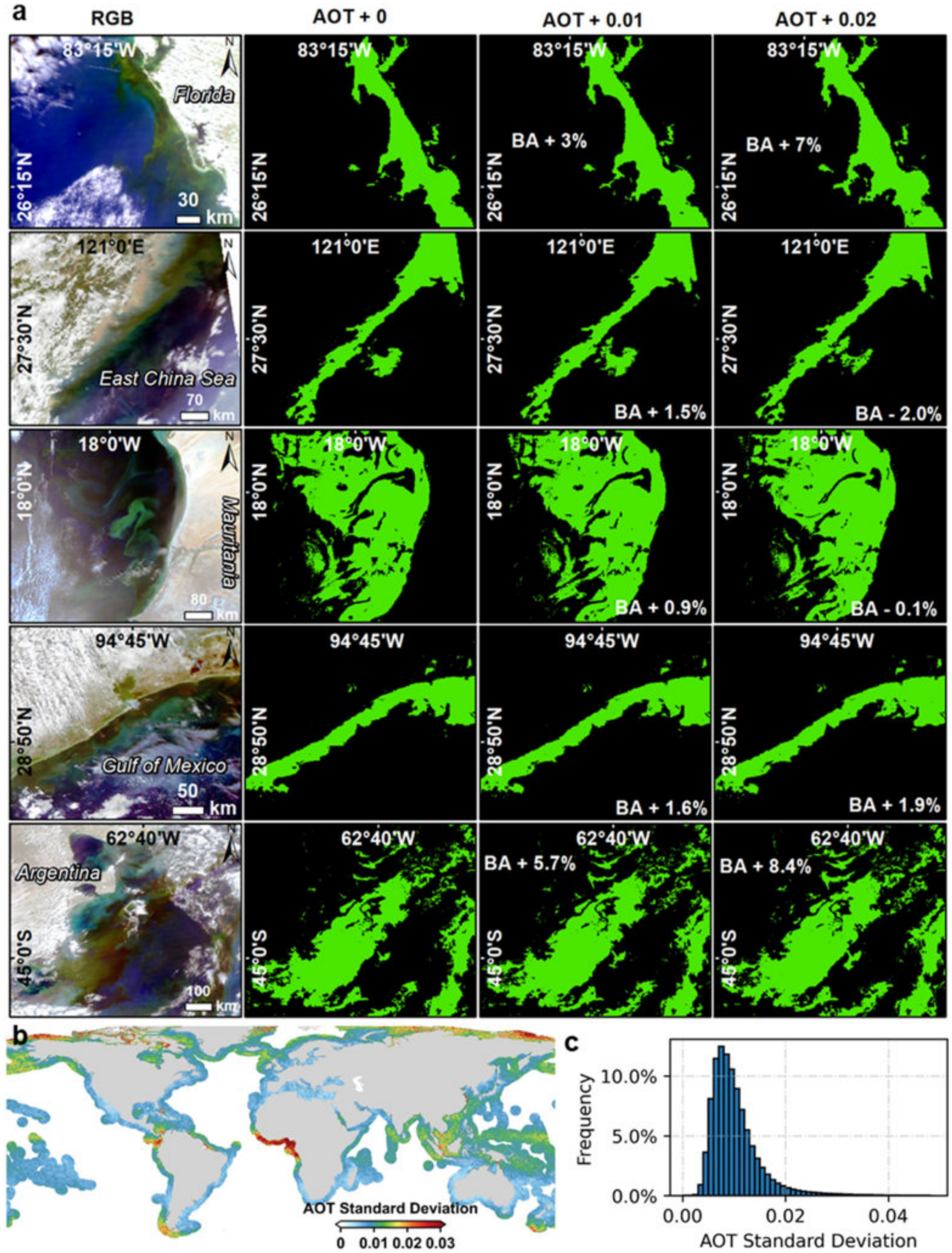
Extended Data Fig. 2 | MODIS-detected bloom count within certain years for several coastal regions with frequently reported blooms. The MODIS observational year is annotated within each panel, and overlaid points indicate *in situ* recorded harmful algal bloom events from the Harmful Algae Event Database (HAEDAT) within the same year. The lower right panel shows the locations of all the HAEDAT records that were used for algorithm validations in this study (Supplementary Table 1), which also demonstrates the increase in sampling effort in the most recent years. Created using ArcMap 10.4.



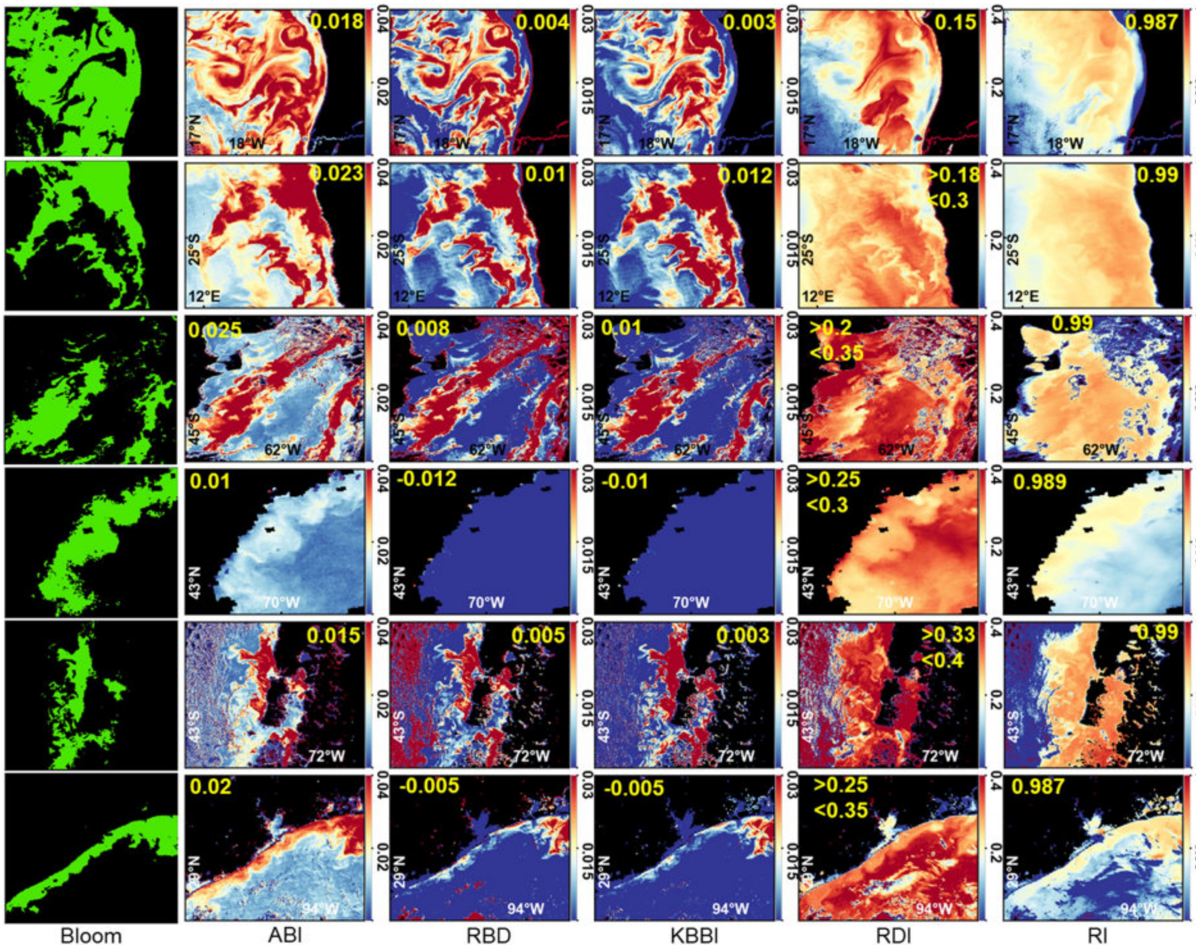
Extended Data Fig. 3 | Performance of the CIE-fluorescence algorithm for phytoplankton bloom detection in 12 selected coastal oceans. From left to right are the RGB-true color composite, ERGB composite, FLHRrc, and the bloom area (green pixels) detected by the CIE-fluorescence algorithm. Created using ArcMap 10.4.



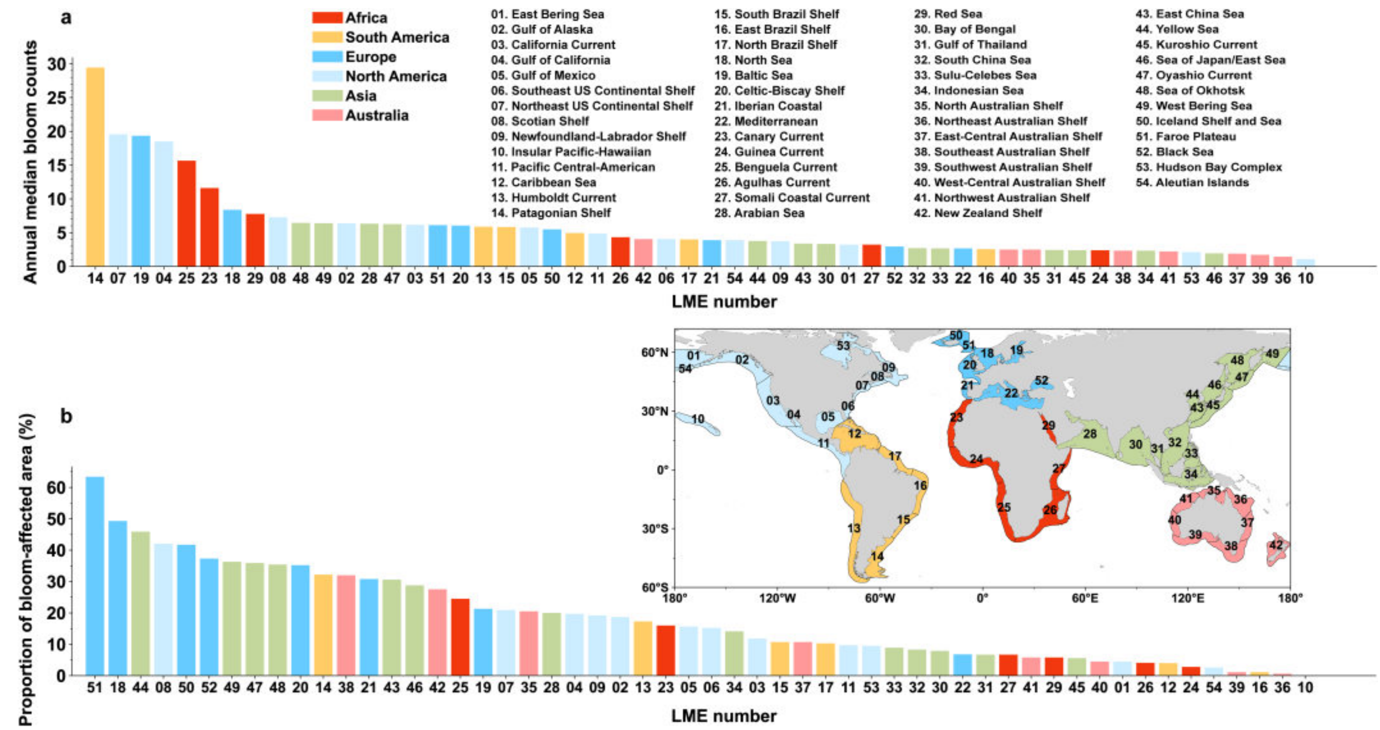
Extended Data Fig. 4 | Examples showing disadvantages of using NASA standard R_{rs} (i.e., with the removal of both Rayleigh and aerosol scattering) in algal bloom detection. From left to right are the RGB composites, ERGB, nFLH, and the bloom areas (green pixels) detected by the CIE-fluorescence algorithm (based on R_{rc}, without the removal of aerosol scattering). Substantial amounts of invalid R_{rs} retrievals can be observed in the red-encircled areas in which severe blooms can be found. Additionally, nFLH shows high values at cloud edges (yellow-encircled areas), making it challenging to use a simple threshold to classify blooms. However, such problems can be circumvented in our CIE-fluorescence algorithm. Created using ArcMap 10.4.



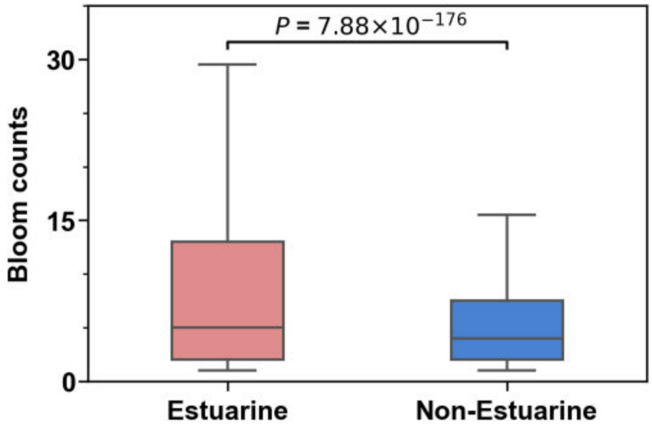
Extended Data Fig. 5 | Sensitivity analysis of the impacts of aerosols on bloom detection. (**a**) Responses of bloom area (BA) to changes in aerosol optical thickness (AOT). Aerosol reflectance (ρ_a) with AOTs of 0.01 and 0.02 at 869-nm is simulated and added to the MODIS images, and the resulting bloom areas (green pixels) with and without added ρ_a are compared. The left columns show the RGB composites, and the right three columns show the bloom areas under different AOTs. The percentages of BA changes are annotated in the panels. (**b**) The standard deviation between the 12 monthly mean values of AOT in global coastal waters (i.e., 66.7% of the intra-annual variability), and the histogram is shown in (**c**). Maps created using ArcMap 10.4.



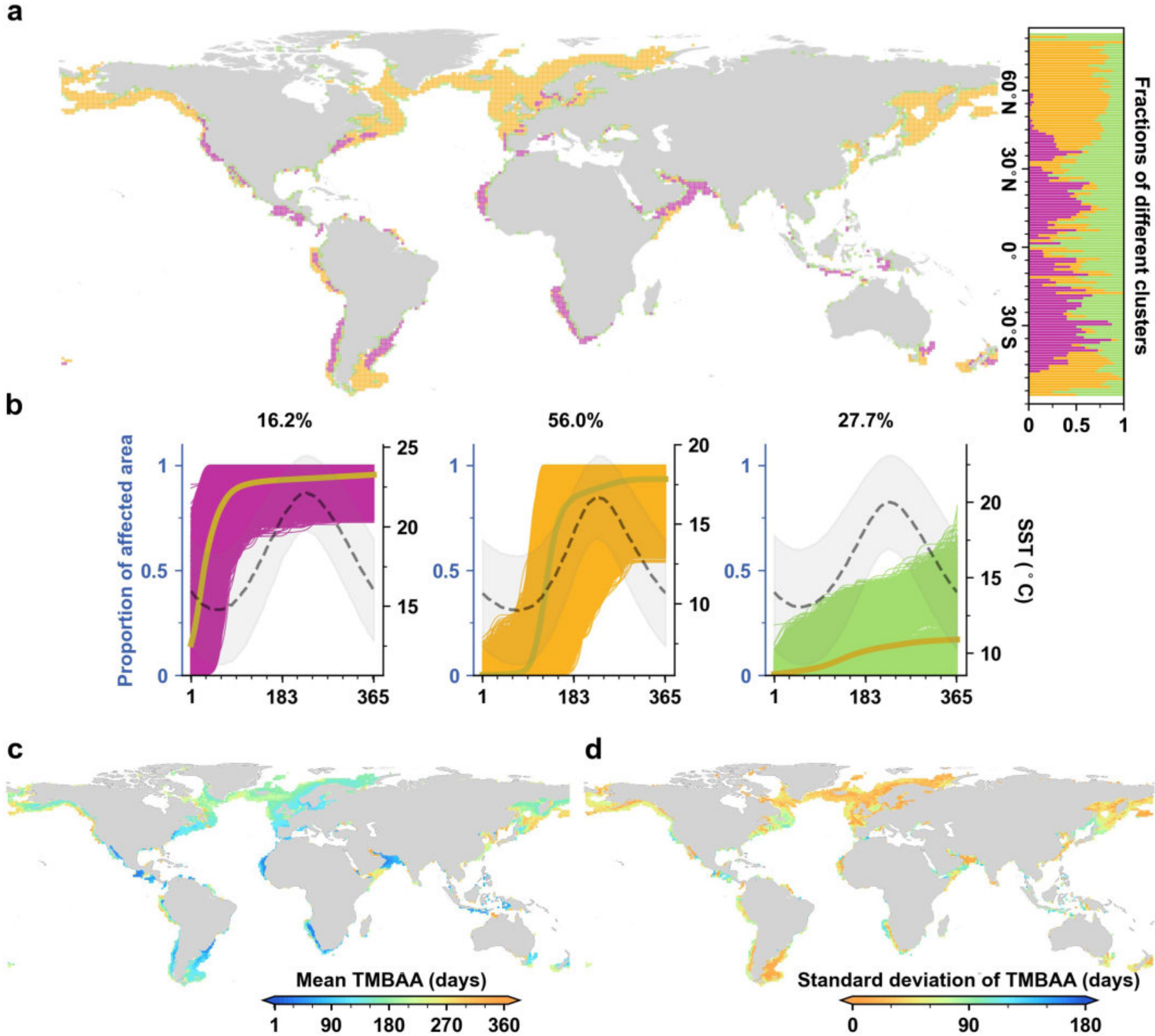
Extended Data Fig. 6 | Comparison of different index-based algorithms in algal bloom detection in various coastal regions. Image-specific thresholds (annotated within the panels) are required (labeled within the panels) for RI[50], ABI (estimated with FLH_{Rrc})[48], RBD[51], KBBI[51], and RDI[52] to delineate accurate bloom areas (i.e., high nFLH values, which appear as bright and darkish features on the ERGB images). The left panels are the bloom areas (green pixels) extracted using our CIE-fluorescence algorithm. The RGB-true color and ERGB composites are shown in Extended Data Fig. 3. Created using ArcMap 10.4.



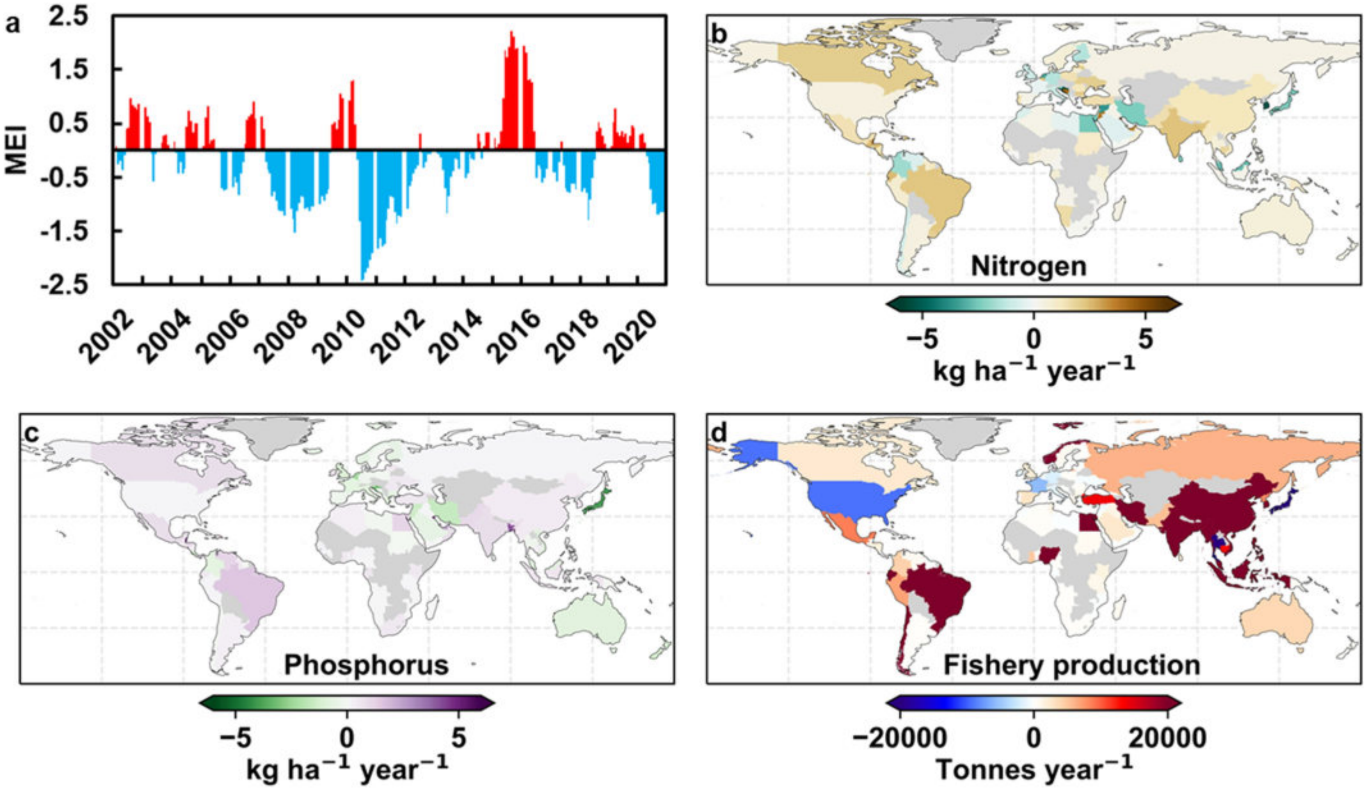
Extended Data Fig. 7 | Annual median bloom count and the proportion of bloom-affected areas for large marine ecosystems (LMEs). (**a**) Annual median bloom count, (**b**) proportion of bloom-affected areas. The data are ordered from the largest to the smallest. The LMEs are grouped by continent, and their names, numbers, and locations are shown in (a) and (b). Map created using Python 3.8.



Extended Data Fig. 8 | Comparison of bloom counts in the estuarine and non-estuarine regions. Boxplots for long-term mean bloom count in the estuarine (n = 13,622 pixel observations) and non-estuarine (n = 361,604 pixel observations) regions. Comparison analysis was performed by two sided Welch's t-test ($P < 0.001$).Upper and lower bounds are first and third quartiles, the bar in the middle represents the median value, and the whiskers show the minimum and maximum values. Sixty-two estuarine zones from large rivers were selected, and the boundary of each zone was manually delineated according to high-resolution satellite images.



Extended Data Fig. 9 | Clusters of different bloom growth paths. (a) The spatial distribution of different clusters. The fractions of different clusters across different latitudes are summarized. (**b**) The development of the maximum bloom-affected areas within a year within 1° × 1° grid cells, where all global grid cells are grouped into three distinct clusters according to the similarity of the bloom growth curve. The colored bond curves represent the mean values of all the grid cells, and their mean SST and associated standard deviations are shown with dashed lines and gray shading. The proportions of different clusters in the global bloom-affected areas are annotated. (**c**) and (**f**) The mean timing of the maximum bloom-affected areas (TMBAA) and the associated standard deviations between 2003 and 2019. The whole year in the Southern Hemisphere is shifted forward by 183 days in (**c**). Maps created using Python 3.8.



Extended Data Fig. 10 | Changes in climate extremes, global fertilizer uses, and fishery production over the past two decades. (a) Changes in the bi-monthly Multivariate El Niño–Southern Oscillation (ENSO) index (MEI) between 2002 and 2020. Positive and negative MEI values represent El Niño and La Niña events, respectively. The dots show annual mean values. **(b–c)** Trends of nitrogen and phosphorus from 2003 to 2019 for different countries. (**d**) Trends of fishery production from 2003 to 2018. Gray indicates no data. Maps created using ArcMap 10.4.

nature portfolio

Corresponding author(s): Lian Feng
Last updated by author(s): Nov 23, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The satellite data were obtained from the U.S. National Aeronautics and Space Administration (NASA) Goddard Space Flight Center (GSFC).
Data analysis	SeaDAS (Version 7.5) were used to analyze the satellite images.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The MODIS Aqua data can be obtained from the U.S. National Aeronautics and Space Administration (NASA) Goddard Space Flight Center (GSFC).
The in situ reported HAB data are available from events from http://haedat.iode.org.
The Exclusive economic zones (EEZs) dataset is available at https://www.marineregions.org/download_file.php?name=World_EEZ_v11_20191118.zip.
The boundaries of large marine ecosystems (LMEs) were obtained from https://www.sciencebase.gov/catalog/item/55c77722e4b08400b1fd8244.

Annual data between 2003 and 2019 on synthetic fertilizer use, including nitrogen and phosphorus, are available from https://ourworldindata.org/fertilizers.
Annual aquaculture production includes cultivated fish and crustaceans in marine and inland waters, and sea tanks, and the data between 2003 and 2018 are available from https://ourworldindata.org/grapher/aquaculture-farmed-fish-production.
The dataset is available from https://psl.noaa.gov/enso/mei/.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design whether sex and/or gender was determined based on self-reporting or assigned and methods used. Provide in the source data disaggregated sex and gender data where this information has been collected, and consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study developed a novel method to map global coastal algal blooms and used this tool to examine satellite images between 2003 and 2020, addressing three fundamental questions: 1) where and how frequently have global coastal oceans been affected by phytoplankton blooms? 2) have the blooms expanded or intensified over the past two decades, both globally and regionally? and 3) what are the potential drivers?
Research sample	Three separate samples were selected. 1) MODIS Aqua images were used to develop the phytoplankton bloom extraction algorithm, 2) MODIS Aqua images and were used to verify the reliability of the algorithm and the accuracy of the phytoplankton bloom extraction results, and 3) in situ reported HAB events from the HAEDAT dataset were used to validate the accuracy of the phytoplankton bloom extraction results.
Sampling strategy	A total of 115 MODIS Aqua images were selected from the different locations where coastal phytoplankton blooms have been recorded in the published literature, of which 80 were used for algorithm development and 35 were used for algorithm validation. A total number of 2609 HAB events that occurred in the coastal area were selected from the HAEDAT dataset.
Data collection	The HAEDAT dataset is a collection of records of harmful algal bloom (HAB) events , maintained under the UNESCO Intergovernmental Oceanographic Commission and with data archives since 1985.
Timing and spatial scale	The satellite data were acquired from different seasons and across various phytoplankton bloom magnitudes between 2003 and 2020, and HAB data from 2003 to 2020 in the HAEDAT dataset were used.
Data exclusions	No data were excluded from analysis.
Reproducibility	Our results could easily be reproduced with existing datasets.
Randomization	Excluding data affected by clouds, a total of 0.76 million MODIS Aqua images from 2003 to 2020 were used to extract phytoplankton blooms in global coastal area.
Blinding	Not applicable in our study.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Article

Evolution of the germline mutation rate across vertebrates

https://doi.org/10.1038/s41586-023-05752-y

Received: 19 November 2021

Accepted: 23 January 2023

Published online: 1 March 2023

Open access

Check for updates

Lucie A. Bergeron[1]✉, Søren Besenbacher[2], Jiao Zheng[3,4], Panyi Li[3], Mads Frost Bertelsen[5], Benoit Quintard[6], Joseph I. Hoffman[7,8], Zhipeng Li[9], Judy St. Leger[10], Changwei Shao[11], Josefin Stiller[1], M. Thomas P. Gilbert[12,13], Mikkel H. Schierup[14] & Guojie Zhang[1,15,16,17]✉

The germline mutation rate determines the pace of genome evolution and is an evolving parameter itself[1]. However, little is known about what determines its evolution, as most studies of mutation rates have focused on single species with different methodologies[2]. Here we quantify germline mutation rates across vertebrates by sequencing and comparing the high-coverage genomes of 151 parent–offspring trios from 68 species of mammals, fishes, birds and reptiles. We show that the per-generation mutation rate varies among species by a factor of 40, with mutation rates being higher for males than for females in mammals and birds, but not in reptiles and fishes. The generation time, age at maturity and species-level fecundity are the key life-history traits affecting this variation among species. Furthermore, species with higher long-term effective population sizes tend to have lower mutation rates per generation, providing support for the drift barrier hypothesis[3]. The exceptionally high yearly mutation rates of domesticated animals, which have been continually selected on fecundity traits including shorter generation times, further support the importance of generation time in the evolution of mutation rates. Overall, our comparative analysis of pedigree-based mutation rates provides ecological insights on the mutation rate evolution in vertebrates.

Germline mutations are the proximate source of genomic innovation and inherited diseases[4]. Consequently, considerable effort has been spent on characterizing the molecular processes underlying these mutations and estimating germline mutation rates (GMRs). Mutations are rare events, yet the frequency at which they are introduced into genomes at each generation varies considerably across taxa, from approximately 10^{-11} mutations per site per generation in unicellular eukaryotes up to approximately 10^{-7} mutations per site per generation in multicellular eukaryotes[1,5,6]. Inferring the driving forces of GMR evolution has important implications for understanding the mechanisms underlying mutagenesis. Several hypotheses have been proposed to explain variation in GMRs among lineages. Some of these invoke molecular mechanisms such as DNA methylation[7] or microsatellite instability[8], whereas others invoke external factors such as exposure to mutagenic environments[9]. Other studies have argued that life-history traits might explain some of the variation both in the prevalence of mutations and in the ability to repair DNA. In particular, the generation time[10] and the metabolic rate[11] have been suggested to be key life-history traits that could be associated with germline mutations. From a long-term evolutionary perspective, the 'drift barrier hypothesis' proposes that lower mutation rates may reflect the increased efficiency of natural selection at reducing the occurrence of mutations in species with large effective population sizes[3].

However, a lack of accurate and standardized GMR estimation has so far precluded testing current hypotheses of GMR evolution. Pedigree-based estimates of GMRs per generation have recently been published for a handful of vertebrate species, mainly focusing on humans and primates[12–17]. Furthermore, a recent comparative study of 16 mammalian species identified an effect of lifespan on somatic mutation rates inferred from the sequencing of intestinal crypts[18]. Nevertheless, interspecific comparisons of GMR variation remain restricted in taxonomic scope[19], partly due to the difficulty of comparing GMR estimates derived using different methodologies[2]. For example, alternative bioinformatic pipelines used in different studies can yield GMR estimates that vary by a factor of two, even when applied to the same parent–offspring trios[2]. This highlights the importance of applying consistent analytical pipelines for interspecies comparisons of GMRs. We therefore generated high-depth genome sequences

[1]Villum Centre for Biodiversity Genomics, Section for Ecology and Evolution, Department of Biology, University of Copenhagen, Copenhagen, Denmark. [2]Department of Molecular Medicine, Aarhus University, Aarhus, Denmark. [3]BGI-Shenzhen, Shenzhen, China. [4]BGI Education Center, University of Chinese Academy of Sciences, Shenzhen, China. [5]Copenhagen Zoo, Frederiksberg, Denmark. [6]Parc Zoologique et Botanique de Mulhouse, Mulhouse, France. [7]Department of Animal Behaviour, Bielefeld University, Bielefeld, Germany. [8]British Antarctic Survey, High Cross, Cambridge, UK. [9]College of Animal Science and Technology, Jilin Agricultural University, Changchun, China. [10]Department of Biomedical Sciences, Cornell University, Ithaca, NY, USA. [11]Key Lab of Sustainable Development of Marine Fisheries, Ministry of Agriculture and Rural Affairs, Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao, China. [12]Center for Evolutionary Hologenomics, The GLOBE Institute, University of Copenhagen, Copenhagen, Denmark. [13]University Museum, NTNU, Trondheim, Norway. [14]Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark. [15]Centre for Evolutionary & Organismal Biology, Women's Hospital, Zhejiang University School of Medicine, Hangzhou, China. [16]Liangzhu Laboratory, Zhejiang University Medical Center, Hangzhou, China. [17]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China. ✉e-mail: lucie.a.bergeron@gmail.com; guojiezhang@zju.edu.cn

(average coverage of more than 67×) for 323 individuals representing 151 trios of 68 vertebrate species, including 36 mammals, 18 birds, 8 ray-finned fishes and 6 reptiles (Supplementary Table 1). We then quantified species-specific GMRs across this wide range of vertebrate taxa using consistent bioinformatics pipelines to test long-standing evolutionary hypotheses on GMR evolution.

Per-generation mutation rate variation

We first estimated the per generation GMR ($\mu_{generation}$) for each trio (that is, mother, father and offspring) by comparing parental and offspring genomes (Fig. 1a, Supplementary Tables 2 and 3 and Supplementary Figs. 1–5 for details on the method). Overall, $\mu_{generation}$ varies by a factor of 40 across all species. On average, mutation rates per generation are higher in reptiles (average of all species 1.17×10^{-8}, 95% CI of the mean = 5.34×10^{-9} to 1.80×10^{-8}) and birds (average of all species 1.01×10^{-8}, 95% CI of the mean = 6.10×10^{-9} to 1.42×10^{-8}) than in mammals (average of all species 7.97×10^{-9}, 95% CI of the mean = 7.04×10^{-9} to 8.90×10^{-9}) and fishes (average of all species 5.97×10^{-9}, 95% CI of the mean = 4.39×10^{-9} to 7.55×10^{-9}). However, the difference among the four major classes of vertebrates is not overall statistically significant (analysis of variance (ANOVA): $F = 1.86$, $P = 0.15$). Furthermore, the amount of variation in $\mu_{generation}$ among species tends to be higher for birds and lower for mammals and fishes (Fig. 1a), although this variation is arguably modest given large differences in life-history traits among these species (for example, there is a 2.8 million-fold difference in the body mass of killer whales and Siamese fighting fish, and there is a 93-fold difference in the generation time between humans and Texas banded geckos).

Species with longer generation intervals are expected to have higher per-generation mutation rates due to a combination of a larger number of cell divisions in spermatogenesis and more time for DNA damage to accumulate[12–14,20]. For the 105 trios for which parental age was known at reproduction, we found a significant positive association between $\mu_{generation}$ and the average parental age at reproduction (linear regression adjusted $r^2 = 0.14$, $P = 3.9 \times 10^{-5}$; Fig. 1b). This pattern is also significant for the 60 mammalian trios with known parental ages (linear regression adjusted $r^2 = 0.37$, $P = 1.6 \times 10^{-7}$) and for the 32 bird trios after excluding a single outlier, the Darwin's rhea (linear regression adjusted $r^2 = 0.31$, $P = 0.0005$). Furthermore, all three of these regressions have similar positive y-intercept values on the order of approximately 0.59×10^{-8} mutations per site per generation. For the trios with known parental ages, paternal and maternal ages at conception are strongly correlated (linear regression adjusted $r^2 = 0.77$, $P < 2.2 \times 10^{-16}$; Extended Data Fig. 1). However, multiple linear regression showed that the age of the father is the most significant explanatory variable (adjusted $r^2 = 0.15$, $P = 9.3 \times 10^{-5}$; paternal age $P = 0.018$; maternal age $P = 0.785$). Thus, a stronger effect of paternal than maternal age on the mutation rate seems to be universal for birds and mammals due to more germline mutations accumulating throughout the life of the male.

The specific types of de novo mutations (DNMs) observed across the 151 trios are concordant with the results of previous studies of individual species[12–14,21–25], including a ratio of transitions over transversions of 2.3 (95% CI on binomial distribution = 2.2–2.5) and a high proportion (48.5%, 95% CI on binomial distribution = 46.7–50.3%) of transitions from strong base pairing to weak base pairing (C:G > T:A) across all DNMs (Supplementary Table 4). Among C:G > T:A mutations, 42.4% (95% CI on binomial distribution = 39.9–45.0%) occurred at CpG sites. The direction of mutations from one base to another (that is, the spectrum of mutation) differed significantly across vertebrate classes ($\chi^2 = 30.0$, d.f. = 15, $P = 0.012$; Supplementary Table 4 and Supplementary Fig. 6). We also found significant differences among vertebrate classes for A > C mutations ($\chi^2 = 16.2$, d.f. = 3, $P = 0.001$) and for C > A mutations ($\chi^2 = 8.8$, d.f. = 3, $P = 0.032$). In particular, fish species exhibit significantly fewer A > C mutations and significantly more C > A mutations than the other vertebrate classes. However, this mutation pattern does not appear to be associated with genome-wide CG content, as overall, the CG content of fishes is similar to that of mammals and birds and lower than that of reptiles (Supplementary Fig. 7). Finally, there is no significant difference between the classes of species in the percentage of all mutations located in CpG sites ($\chi^2 = 4.3$, d.f. = 3, $P = 0.23$), implying that high mutation rates at CpG sites are a conserved feature across vertebrates.

Variable male-driven evolution

In mammals and birds, the much larger number of germ-cell divisions per generation in the male germ line leads to the expectation of a male mutation rate bias, coined the 'male-driven evolution hypothesis'[26,27]. However, very little is known about interspecific variation in the magnitude of the male-to-female ratio of the contribution of germline mutations (α). Previous studies have reported high α values in mammals (ranging from 1.0 to 20.1)[28] and birds (ranging from 3.9 to 6.5)[29] based on indirect estimates obtained by comparing rates of sequence divergence on the autosomes and sex chromosomes (see Extended Data Fig. 2 and Supplementary Table 5). However, other evolutionary forces can also act differently on the X chromosome and autosomes. For example, stronger natural selection on the X chromosome could lead to lower than expected divergence from the common ancestor, upwardly biasing estimates of α[28]. Furthermore, estimates of α derived in this way are averages over a phylogenetic branch and may thus differ from the contemporary species α. Here we directly quantified α by assigning the parental origin of the DNMs. Around 48% of all 3,034 DNMs across all of the trios could be phased to their parental origin (see Supplementary Table 6 for positions of all mutations). Owing to the relatively small number of mutations in each trio (Supplementary Table 2), we analysed male bias after taxonomically grouping the species into classes and orders (Fig. 1c).

Mammals showed a male bias of α = 2.3 (95% CI = 2.0–2.6). In general, our α estimates are in line with previous estimates derived for similar species based on genome alignments[30,31]. For example, we found that among mammals, primates have the largest male bias with α = 3.8 (95% CI = 2.6–5.7), similar to what was previously reported for several species belonging to this group[12–14,21,22,32,33]. Rodents have the lowest male bias among the mammals in our study, with α = 2.1 (95% CI = 1.4–3.1), consistent with a previous study based on mouse pedigrees[34]. This pattern can be explained by the short generation time of rodents, which leads to a smaller difference in cell divisions between the male and female germ lines[35]. However, the variation in α is relatively small given the variation in generation time among species (for example, between 30 years for humans and 8 months for the short-tailed opossum). Thus, an alternative hypothesis to explain the observed α would be a higher contribution of DNA damage, specifically in the male germ line for species with large generation times[31].

Birds also showed an overall high male bias with α = 3.2 (95% CI = 2.5–4.1), although there is appreciable variation among different lineages. In particular, passerine birds and waterbirds (Pelecaniformes and Sphenisciformes) exhibited the largest male bias, both with α = 7.6 (95% CI = 4.3–13.5 for Passeriformes and 95% CI = 3.5–16.3 for Pelecaniformes and Sphenisciformes). High levels of male–male competition will lead to an increased amount of sperm being produced and faster sperm turnover, which would be expected to cause a higher male bias[36]. Indeed, many passerine birds have large cloacal protuberances[37] and relatively heavy testes[38], which are often used as proxies of sperm competition[39]. For instance, in two of the passerine species included in our study, testes represent between 1.2% (for *Turdus merula*) and over 2% (for *Saxicola maurus*) of the total body mass[38]. Moreover, extra-pair mating is common in many passerine birds[40] as well as in penguins[41], also indicating a high level of sperm competition. Overall, our results lend further support to the male-driven hypothesis in birds and mammals[27].

By contrast, reptiles have a relatively small male bias with α = 1.5 (95% CI = 1.2–1.8), whereas fishes appear to have a greater proportion of

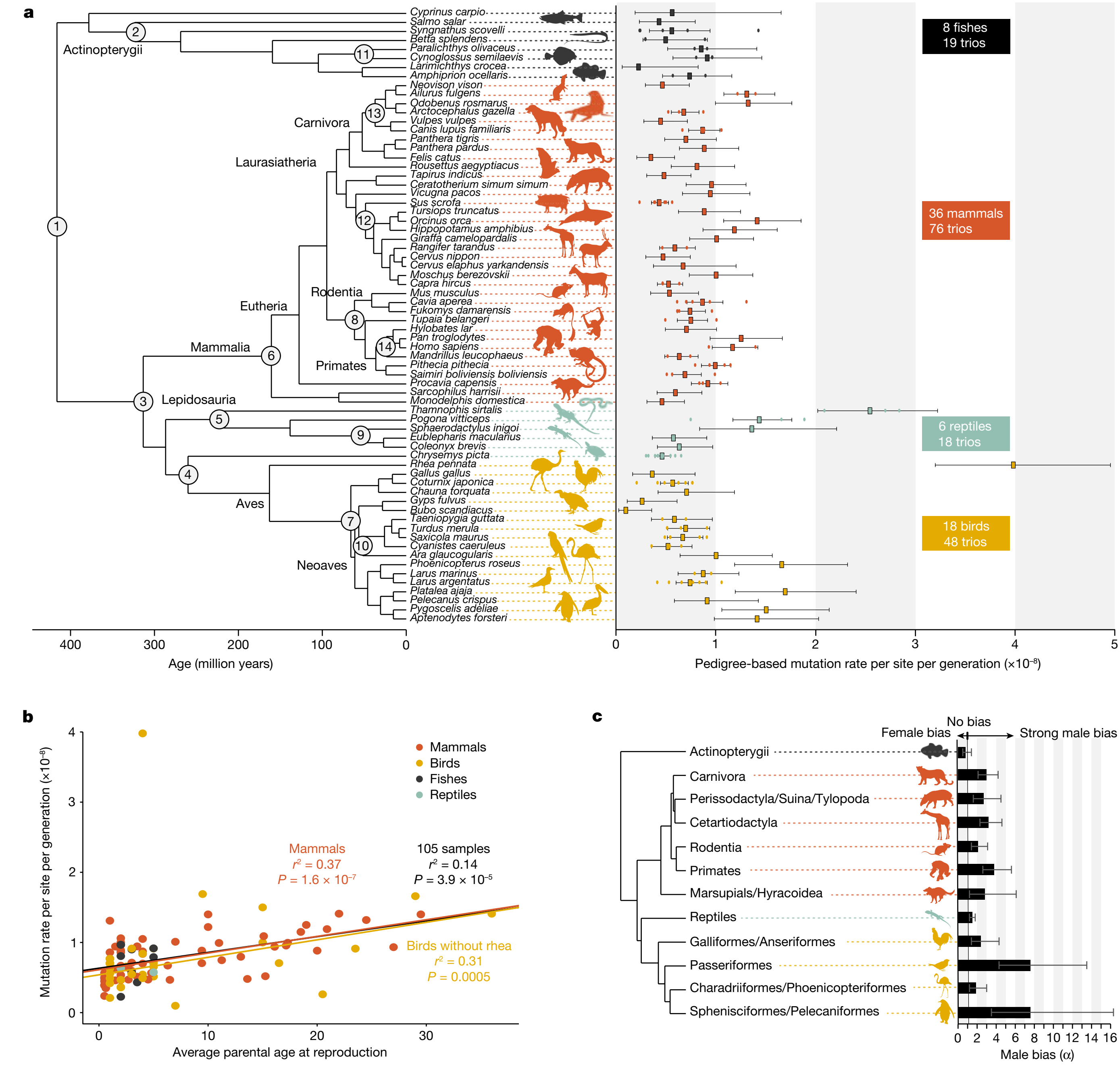


Fig. 1 | Variation in GMRs and their association with life-history traits across 68 vertebrate species. a, The phylogenetic tree of 68 species is based on UCE data and is calibrated with fossil data at 14 nodes (see Methods; Extended Data Fig. 3 and Supplementary Fig. 8). The average pedigree-based mutation rates per generation for each species, which are represented by the squares, show 40-fold variation among species. The 95% binomial confidence intervals are shown, and individual trios are represented by round points. See Supplementary Table 8 and Extended Data Fig. 4 for a comparison with published estimated rates of closely related species. **b**, The per-generation mutation rate is significantly associated with the average parental age at the time of offspring production across all individuals with known paternal age (105 trios), using linear regression. For birds, this relationship is statistically significant after removing a single outlier, the Darwin's rhea. **c**, The male-to-female contribution ratio (α) is estimated for groups of vertebrates having at least 30 mutations phased to their parents of origin in each group. The highest male bias (7.6:1) is found in two bird lineages, whereas fishes and reptiles show negligible male bias. The data are represented with 95% confidence intervals based on the binomial variance. The silhouette of *Syngnathus scovelli* was created by J.S. All other silhouettes are from PhyloPic (http://phylopic.org), except for one of the silhouettes of *Sarcophilus harrisii*, which was created by S. Werning, and the silhouette of *Pan troglodytes*, which was created by T. M. Keesey (vectorization) and T. Hisgett (photography); both are available under a CC-BY 3.0 licence (https://creativecommons.org/licenses/by/3.0).

mutations of maternal origin (α = 0.8), although the 95% CI overlaps 1 (95% CI = 0.5–1.4). This variation among vertebrate classes can be explained by differences in the process of gametogenesis. Although most birds and mammals produce sperm cells continuously through time[42], reptiles and fishes tend to be seasonal breeders, producing sperm cells during a limited period before the mating season[43–45], which will tend to reduce differences in cell division numbers between males and females, leading to more equal α. Moreover, female fishes are usually synchronous ovulators[46], producing hundreds to millions of eggs at the same time followed by a proliferation of new oogonia[47]. This

implies that females continually produce germ cells throughout their life, which would further reduce the difference in cell division number between males and females.

Species with lower sex bias also exhibited a larger proportion of shared mutations between siblings, with 12.0% (s.e. of 6.5%) of shared mutations between siblings for fish and 8.1% (s.e. of 5.3%) for reptiles compared with 1.5% (s.e. of 0.7%) for mammals and 2.2% (s.e. of 1.4%) for birds (Supplementary Table 7). An explanation for the repeated occurrence of those mutations is that they appear during the primordial germ cell specification in one of the parents[48]. The occurrence of primordial germ cell specification mutations is independent of parental sex. Consequently, a higher number of primordial germ cell specification mutations in some vertebrate groups could be an alternative explanation for the lower male-biased contribution to DNMs.

Yearly mutation rates

To use our results for phylogenetic dating and to compare the speed of evolution among species with different generation times, we needed estimates of yearly mutation rates. Different methods have been used in the literature to estimate yearly mutation rates. When sample sizes are small, yearly rates are commonly inferred by dividing the per-generation rate by the average age of the parents (or the generation time if parental age is unknown)[49–51]. However, this method implicitly assumes a constant accumulation of mutations from conception to reproduction, that is, the regression line of mutation rate on parental age should run through the origin. Our results (Fig. 1b), as well as previous studies of mice, humans and cats[20,34], imply that parents always carry a minimum number of mutations in their gametes regardless of their age. This could lead to the yearly rate being overestimated for a given species if the sampled trio (or trios) had young parents compared with the average generation time for that species[52]. Consequently, we built a model that incorporates this mutational contribution at birth. Unfortunately, small per-species sample sizes in our dataset precluded modelling the effects of parental age separately for each species. However, we observed very similar intercepts and slopes across taxonomic groups, allowing us to fit a joint model for all species. A Poisson model explaining the number of mutations in each trio using a mutational contribution at birth and a weighted average of paternal and maternal age fits the data surprisingly well. To incorporate interspecific variation in male bias, we used the per-species fraction of paternal and maternal mutations estimated using read-backed phasing to weigh the average of the parental ages for each trio. Using this model, the number of predicted mutations matches the observed number with an overall r^2 of 0.73 (mammalian $r^2 = 0.58$, avian $r^2 = 0.51$; Supplementary Note 1).

The yearly rates inferred with the naive method of dividing the per-generation rate by parental age (μ_{yearly}) and the rates obtained with our model ($\mu_{yearly_modelled}$) yielded similar results (Pearson's correlation $r^2 = 0.40$, $P = 0.002$), and for 55% of the species, μ_{yearly} falls within the 95% confidence interval of the $\mu_{yearly_modelled}$. As expected, the estimates showed the greatest differences for those species in which the parents reproduced far from the generation time, with the model-based estimates being smaller for those species that reproduced earlier than their generation time and larger for those species that reproduced later than their generation time. For example, the pigs in our dataset reproduced at around 6 months of age, which is more than 5 years earlier than the estimated generation time of this species. Thus, $\mu_{yearly} = 8.64 \times 10^{-9}$ mutations per site per year was potentially overestimated compared with the $\mu_{yearly_modelled} = 1.05 \times 10^{-9}$ mutations per site per year at the generation time. Conversely, the yearly rate of the Texas banded gecko was potentially underestimated at $\mu_{yearly} = 3.17 \times 10^{-9}$ mutations per site per year using the reproductive age of 2 years of age from our dataset, whereas the modelled rate was $\mu_{yearly_modelled} = 1.96 \times 10^{-8}$ mutations per site per year at a generation time of between 3 and 4 months. Both the naive method and the modelled method have been used in the literature

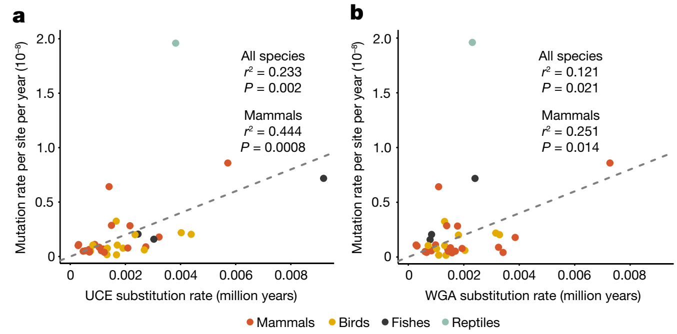


Fig. 2 | GMRs are associated with long-term substitution rates. a,**b**, There is a positive association between the modelled yearly pedigree-based mutation rates and the macroevolutionary substitution rates when using phylogenetic regression (PGLS) on both UCEs and their flanking sequences (**a**) and whole-genome alignments (WGAs) (**b**). The grey dashed lines indicate equality. See Extended Data Fig. 5 for plots of the same data on a log scale and Extended Data Fig. 6 for a comparison of UCE and WGA methods.

to estimate yearly rates and both have caveats owing to the underlying assumptions they require. Bearing this in mind, we decided to use $\mu_{yearly_modelled}$ for the current analysis as we believe that this measure is more representative of the yearly rate at the generation time for each species (estimated yearly rates are provided in Supplementary Table 9 for comparison).

The estimated average $\mu_{yearly_modelled}$ varies more than 120-fold among species (Supplementary Note 1 and Supplementary Table 9), with the highest $\mu_{yearly_modelled}$ estimated for the Texas banded gecko at 1.96×10^{-8} mutations per site per year (95% CI = 1.23×10^{-8} to 2.83×10^{-8}), whereas the lowest $\mu_{yearly_modelled}$ estimates were obtained for two bird species, the griffon vulture and the snowy owl, both with less than 0.18×10^{-9} mutations per site per year (snowy owl: $\mu_{yearly_modelled} = 0.16 \times 10^{-9}$, 95% CI = 0.05×10^{-9} to 0.34×10^{-9}; griffon vulture: $\mu_{yearly_modelled} = 0.17 \times 10^{-9}$, 95% CI = 0.07×10^{-9} to 0.32×10^{-9}). This large amount of interspecific variation is remarkable given that pedigree-based GMR estimates of individual species assessed by previous separate studies only show an approximately 16-fold variation in yearly GMRs[34,51]. Within primates, we observed a twofold variation across species and found a general trend for rates to be higher in the New World monkeys than in the great apes. This is consistent with previous independent estimates from primates[19] and supports the 'hominoid slowdown' hypothesis[53–56].

Next, we used $\mu_{yearly_modelled}$ to assess the strength of the association between GMRs and long-term evolutionary substitution rates. To obtain an estimate of the long-term substitution rate, we used the alignment of ultraconserved elements (UCEs), which are more likely to align among taxonomically distant species, plus 1,000 bp of flanking regions on each side of the UCE sequences, which will more closely reflect the neutral substitution rate[57]. We found a significant positive correlation between $\mu_{yearly_modelled}$ and the UCE substitution rate after excluding domesticated species owing to their overall much higher yearly mutation rates (see the following section; phylogenetic generalized least squares (PGLS): adjusted $r^2 = 0.23$, $P = 0.002$; Fig. 2a). This pattern is especially pronounced for mammals (PGLS: adjusted $r^2 = 0.44$, $P = 0.0008$), even after removing the two outliers (PGLS: adjusted $r^2 = 0.32$, $P = 0.009$). We also found a significant relationship between $\mu_{yearly_modelled}$ and the long-term substitution rate inferred using whole-genome alignments (PGLS: adjusted $r^2 = 0.12$, $P = 0.02$; Fig. 2b).

Life-history traits shape GMR variation

To test various hypotheses relating to the causes of GMR variation among species, we tested for associations between the modelled mutation rate per generation ($\mu_{generation_modelled}$) and life-history traits

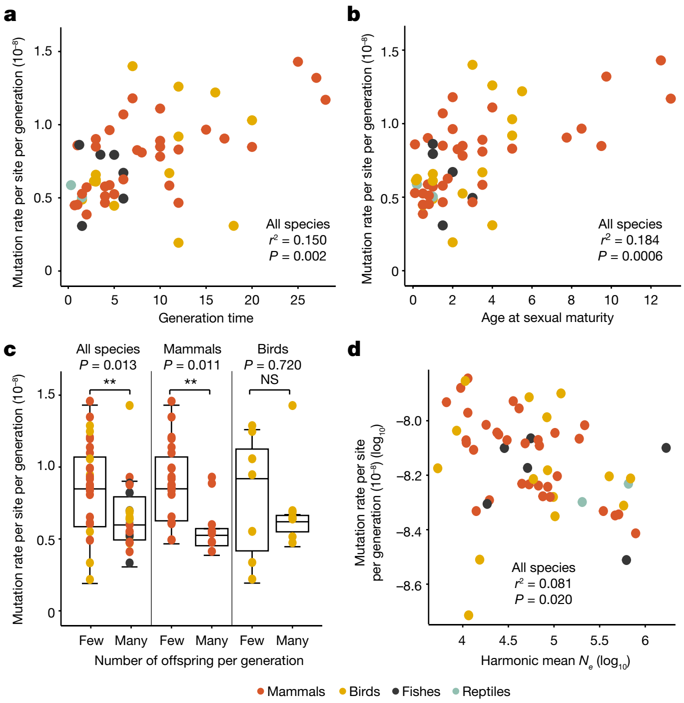


Fig. 3 | Predictors of interspecific variation in GMRs. a–**c**, Significant positive associations are found using phylogenetic regression (PGLS) between the modelled per-generation mutation rates and three life-history traits: species-specific mean generation time (**a**), age at sexual maturity (**b**) and the number of offspring per generation (**c**). In total there are 55 species with modelled per-generation rates, including 32 mammalian and 15 avian species. The box plot in **c** represents the median, the interquartile range and the maximum and minimum excluding outliers. **d**, Species-specific average per-generation mutation rates are negatively associated with the harmonic mean of the effective population size (N_e) over the past 1 million years, using phylogenetic regression (PGLS).

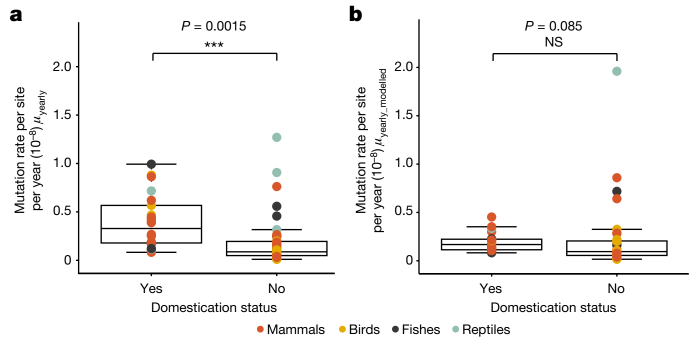


Fig. 4 | The yearly GMRs are higher in domesticated species than in non-domesticated species. a, Yearly GMRs are significantly higher in domesticated or farmed species than in wild species (using phylogenetic regression (PGLS) on a total of 68 species). **b**, Using the modelled mutation rate instead (using phylogenetic regression (PGLS) on a total number of 55 species) shows that there is no difference in yearly GMRs between domesticated and non-domesticated animals, suggesting that this difference is mainly driven by the shorter generation time of domesticated species. The box plots represent the median, the interquartile range and the maximum and minimum excluding outliers.

including mating system (monogamy versus polygamy), maturation time, body mass, longevity, fecundity and the generation time (Supplementary Table 9). We used the $\mu_{generation_modelled}$ instead of the $\mu_{generation}$ as the former is less dependent on the age of the parents and is more representative of the rate at generation time for a given species. Although taking into account phylogenetic relatedness, many of these traits are significantly associated with $\mu_{generation_modelled}$ including the generation time (PGLS: adjusted $r^2 = 0.15$, $P = 0.002$; Fig. 3a), the maturation time (PGLS: adjusted $r^2 = 0.18$, $P = 0.0006$; Fig. 3b) and the number of offspring per generation (PGLS: adjusted $r^2 = 0.10$, $P = 0.013$; Fig. 3c). Species with a higher number of offspring per generation also showed significantly lower $\mu_{generation_modelled}$ when considering only mammalian species (PGLS: adjusted $r^2 = 0.17$, $P = 0.011$), but this relationship was not significant for birds (PGLS: adjusted $r^2 = -0.066$, $P = 0.720$). Collectively, these traits explain almost 18% of the variation in $\mu_{generation_modelled}$ (multiple PGLS: adjusted $r^2 = 0.18$, $P = 0.004$). The other life-history traits that we tested, including longevity, mating strategy and body mass, are not significantly associated with $\mu_{generation_modelled}$ (see Extended Data Fig. 7).

Another key parameter for species evolution is the effective population size (N_e), which impacts genetic drift and the efficacy of selection. To investigate the effect of N_e on $\mu_{generation_modelled}$ and to test the drift barrier hypothesis[3], which predicts the evolution of higher mutation rates in species with small N_e, we calculated N_e using the pairwise sequentially Markovian coalescent method based on one randomly selected father per species. To avoid circularity, we estimated N_e based on the substitution rate calculated from the UCE alignment (Supplementary Table 9). Indeed, if N_e was estimated using the pedigree-based mutation rate, a stronger correlation might arise between N_e and the mutation rate (see Extended Data Fig. 8). We found a significant negative association between $\mu_{generation_modelled}$ and the harmonic mean N_e per species over the past 30,000–1,000,000 years (PGLS: adjusted $r^2 = 0.08$, $P = 0.020$; Fig. 3d) as would be expected under the drift barrier hypothesis. This relationship is mainly driven by mammals (PGLS: adjusted $r^2 = 0.31$, $P = 0.0006$), a signal that is also observed when using the harmonic average N_e over a smaller timescale (30,000–130,000 years; PGLS: adjusted $r^2 = 0.10$, $P = 0.04$, Extended Data Fig. 8). The most appropriate timeframe used to estimate N_e depends on the evolutionary time necessary for the mutation rate to adapt to changes in N_e. However, the pairwise sequentially Markovian coalescent method cannot accurately estimate recent N_e. To overcome this limitation, we also estimated N_e as $\pi/4\mu$, with nucleotide diversity (π) and the substitution rate per site per generation (μ) estimated from the UCE alignments. This results in a similar negative association between N_e and $\mu_{generation_modelled}$ (linear regression: adjusted $r^2 = 0.83$, $P = 2.2 \times 10^{-16}$; Extended Data Fig. 9), further supporting the drift barrier hypothesis. However, caution should be taken as N_e estimates rely on generation times inferred from contemporary observations, whereas generation times could conceivably have changed over evolutionary timescales. Furthermore, population size depends negatively on the generation time (PGLS N_e in log scale: adjusted $r^2 = 0.20$, $P = 0.0004$). Therefore, a negative association between N_e and μ could potentially be driven by a large effect of the generation time on per-generation mutation rates.

High yearly rates in domesticated species

Domestication imposes strong artificial selection, recurrent genetic bottlenecks or both. Our dataset includes 22 domesticated or semi-wild species that have been bred in captivity for many generations. When using the naive method of dividing the per-generation rate by the parental age, these species show significantly higher μ_{yearly} than the non-domesticated species (PGLS: adjusted $r^2 = 0.13$, $P = 0.0015$; Fig. 4a). The higher mutation rates of domesticated animals are likely due to strong artificial selection for traits such as shorter generation times. Indeed, using $\mu_{yearly_modelled}$, we found no difference between domesticated and non-domesticated species (PGLS: adjusted $r^2 = 0.037$,

P = 0.08; Fig. 4b). Consequently, the higher yearly mutation rate observed in domesticated species is more likely to be explained by the lowering of reproductive age associated with domestication rather than by an inherent change to the mutational process caused by relaxed selection on the mutation rate due to small population sizes and bottlenecks associated with domestication[58,59].

Conclusions

Here we analysed pedigree-based GMR variation in an unprecedentedly broad phylogenetic context. We showed that there is a consistent male bias in mammals and birds, whereas reptiles and fish exhibited more evenly matched contributions of DNMs between parents. This could be due to contrasting mutagenic processes, such as differences in male and female germline cell division observed in mammals and birds, or differences among species in the proportion of DNMs occurring in primordial germ cell specification versus in the parental germ lines. Our results also support the drift barrier hypothesis, as we found a negative association between the per-generation mutation rate and effective population size. Moreover, our results suggest that an appreciable proportion of the variation in the GMR can be explained by life-history traits, including maturation time and the number of offspring per generation. Our study also highlights the importance of the generation time, as illustrated by the particular case of domesticated animals, in which exceptionally high yearly mutation rate estimates can be explained by artificially induced short generation times. In addition, some of the trio samples in our study were collected from captive animals at zoos or conservation centres. These populations might have different generation times than those in the wild, which could potentially introduce biases into some of our mutation rate estimates. Future studies should focus on wild pedigree samples, which can be accessed from long-term conservation and monitoring programmes[60].

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-05752-y.

1. Lynch, M. et al. Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.* **17**, 704–714 (2016).
2. Bergeron, L. A. et al. The mutationathon highlights the importance of reaching standardization in estimates of pedigree-based germline mutation rates. *eLife* **11**, e73577 (2022).
3. Lynch, M. Evolution of the mutation rate. *Trends Genet.* **26**, 345–352 (2010).
4. Acuna-Hidalgo, R., Veltman, J. A. & Hoischen, A. New insights into the generation and role of de novo mutations in health and disease. *Genome Biol.* **17**, 241 (2016).
5. Sturtevant, A. H. Essays on evolution. I. On the effects of selection on mutation rate. *Q. Rev. Biol.* **12**, 464–467 (1937).
6. Zhang, G. The mutation rate as an evolving trait. *Nat. Rev. Genet.* **24**, 3 (2022).
7. Mugal, C. F., Arndt, P. F., Holm, L. & Ellegren, H. Evolutionary consequences of DNA methylation on the GC content in vertebrate genomes. *G3* **5**, 441–447 (2015).
8. Baer, C. F., Miyamoto, M. M. & Denver, D. R. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nat. Rev. Genet.* **8**, 619–631 (2007).
9. Wright, S. D., Ross, H. A., Jeanette Keeling, D., McBride, P. & Gillman, L. N. Thermal energy and the rate of genetic evolution in marine fishes. *Evol. Ecol.* **25**, 525–530 (2011).
10. Ohta, T. An examination of the generation-time effect on molecular evolution. *Proc. Natl Acad. Sci. USA* **90**, 10676–10680 (1993).
11. Martin, A. P. & Palumbi, S. R. Body size, metabolic rate, generation time, and the molecular clock. *Proc. Natl Acad. Sci. USA* **90**, 4087–4091 (1993).
12. Bergeron, L. A. et al. The germline mutational process in rhesus macaque and its implications for phylogenetic dating. *Gigascience* **10**, giab029 (2021).
13. Wu, F. L. et al. A comparison of humans and baboons suggests germline mutation rates do not track cell divisions. *PLoS Biol.* **18**, e3000838 (2020).
14. Wang, R. J. et al. Paternal age in rhesus macaques is positively associated with germline mutation accumulation but not with measures of offspring sociability. *Genome Res.* **30**, 826–834 (2020).
15. Campbell, C. R. et al. Pedigree-based and phylogenetic methods support surprising patterns of mutation rate and spectrum in the gray mouse lemur. *Heredity* **127**, 233–244 (2021).
16. Besenbacher, S., Hvilsom, C., Marques-Bonet, T., Mailund, T. & Schierup, M. H. Direct estimation of mutations in great apes reconciles phylogenetic dating. *Nat. Ecol. Evol.* **3**, 286–292 (2019).
17. Thomas, G. W. C. et al. Reproductive longevity predicts mutation rates in primates. *Curr. Biol.* **28**, 3193–3197.e5 (2018).
18. Cagan, A. et al. Somatic mutation rates scale with lifespan across mammals. *Nature* **604**, 517–524 (2022).
19. Chintalapati, M. & Moorjani, P. Evolution of the mutation rate across primates. *Curr. Opin. Genet. Dev.* **62**, 58–64 (2020).
20. Wang, R. J. et al. De novo mutations in domestic cat are consistent with an effect of reproductive longevity on both the rate and spectrum of mutations. *Mol. Biol. Evol.* **39**, msac147 (2022).
21. Venn, O. et al. Strong male bias drives germline mutation in chimpanzees. *Science* **344**, 1272–1275 (2014).
22. Jónsson, H. et al. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017).
23. Tatsumoto, S. et al. Direct estimation of de novo mutation rates in a chimpanzee parent-offspring trio by ultra-deep whole genome sequencing. *Sci. Rep.* **7**, 13561 (2017).
24. Yuen, R. K. C. et al. Genome-wide characteristics of de novo mutations in autism. *npj Genomic Med.* **1**, 160271–1602710 (2016).
25. Wang, H. & Zhu, X. De novo mutations discovered in 8 Mexican American families through whole genome sequencing. *BMC Proc.* **8**, S24 (2014).
26. Li, W.-H., Yi, S. & Makova, K. Male-driven evolution. *Curr. Opin. Genet. Dev.* **12**, 650–656 (2002).
27. Miyata, T., Hayashida, H., Kuma, K., Mitsuyasu, K. & Yasunaga, T. Male-driven molecular evolution: a model and nucleotide sequence analysis. *Cold Spring Harb. Symp. Quant. Biol.* **52**, 863–867 (1987).
28. Wilson Sayres, M. A. & Makova, K. D. Genome analyses substantiate male mutation bias in many species. *BioEssays* **33**, 938–945 (2011).
29. Ellegren, H. & Fridolfsson, A.-K. Male-driven evolution of DNA sequences in birds. *Nat. Genet.* **17**, 182–184 (1997).
30. Sayres, M. A. W., Venditti, C., Pagel, M. & Makova, K. D. Do variations in substitution rates and male mutations bias correlate with life-history traits? A study of 32 mammalian genomes. *Evolution* **65**, 2800–2815 (2011).
31. de Manuel, M., Wu, F. L. & Przeworski, M. A paternal bias in germline mutation is widespread in amniotes and can arise independently of cell division numbers. *eLife* **11**, e80008 (2022).
32. Francioli, L. C. et al. Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.* **47**, 822–826 (2015).
33. Gao, Z. et al. Overlooked roles of DNA damage and maternal age in generating human germline mutations. *Proc. Natl Acad. Sci. USA* **116**, 9491–9500 (2019).
34. Lindsay, S. J., Rahbari, R., Kaplanis, J., Keane, T. & Hurles, M. E. Similarities and differences in patterns of germline mutation between mice and humans. *Nat. Commun.* **10**, 4053 (2019).
35. Gibbs, R. A. et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–520 (2004).
36. Blumenstiel, J. P. Sperm competition can drive a male-biased mutation rate. *J. Theor. Biol.* **249**, 624–632 (2007).
37. Birkhead, T. R., Briskie, J. V. & Møller, A. P. Male sperm reserves and copulation frequency in birds. *Behav. Ecol. Sociobiol.* **32**, 85–93 (1993).
38. Moller, A. P. Sperm competition, sperm depletion, paternal care, and relative testis size in birds. *Am. Nat.* **137**, 882–906 (1991).
39. Birkhead, T. R. & Montgomerie, R. Three decades of sperm competition in birds. *Phil. Trans. R. Soc. B* **375**, 20200208 (2020).
40. Brouwer, L. & Griffith, S. C. Extra-pair paternity in birds. *Mol. Ecol.* **28**, 4864–4882 (2019).
41. Hunter, F. M., Harcourt, R., Wright, M. & Davis, L. S. Strategic allocation of ejaculates by male Adélie penguins. *Proc. R. Soc. Lond. B* **267**, 1541–1545 (2000).
42. Hamamah, S. & Gatti, J. L. Role of the ionic environment and internal pH on sperm activity. *Hum. Reprod.* **13**, 20–30 (1998).
43. Gribbins, K. Reptilian spermatogenesis. *Spermatogenesis* **1**, 250–269 (2011).
44. Gribbins, K. M., Gist, D. H. & Congdon, J. D. Cytological evaluation of spermatogenesis and organization of the germinal epithelium in the male slider turtle, *Trachemys scripta*. *J. Morphol.* **255**, 337–346 (2003).
45. Schulz, R. W. et al. Spermatogenesis in fish. *Gen. Comp. Endocrinol.* **165**, 390–411 (2010).
46. Lubzens, E., Young, G., Bobe, J. & Cerdà, J. Oogenesis in teleosts: how fish eggs are formed. *Gen. Comp. Endocrinol.* **165**, 367–389 (2010).
47. Jalabert, B. Particularities of reproduction and oogenesis in teleost fish compared to mammals. *Reprod. Nutr. Dev.* **45**, 261–279 (2005).
48. Jónsson, H. et al. Multiple transmissions of de novo mutations in families. *Nat. Genet.* **50**, 1674–1680 (2018).
49. Martin, H. C. et al. Insights into platypus population structure and history from whole-genome sequencing. *Mol. Biol. Evol.* **35**, 1238–1252 (2018).
50. Smeds, L., Qvarnström, A. & Ellegren, H. Direct estimate of the rate of germline mutation in a bird. *Genome Res.* **26**, 1211–1218 (2016).
51. Feng, C. et al. Moderate nucleotide diversity in the Atlantic herring is associated with a low mutation rate. *eLife* **6**, e23907 (2017).
52. Gao, Z., Wyman, M. J., Sella, G. & Przeworski, M. Interpreting the dependence of mutation rates on age and time. *PLoS Biol.* **14**, e1002355 (2016).
53. Goodman, M. Rates of molecular evolution: the hominoid slowdown. *BioEssays* **3**, 9–14 (1985).
54. Moorjani, P., Amorim, C. E. G., Arndt, P. F. & Przeworski, M. Variation in the molecular clock of primates. *Proc. Natl Acad. Sci. USA* **113**, 10607–10612 (2016).
55. Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.* **13**, 745–753 (2012).
56. Soojin, V. Y. Morris Goodman's hominoid rate slowdown: the importance of being neutral. *Mol. Phylogenet. Evol.* **66**, 569–574 (2013).

57. Faircloth, B. C. et al. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol.* **61**, 717–726 (2012).
58. Garcia, J. A. & Lohmueller, K. E. Negative linkage disequilibrium between amino acid changing variants reveals interference among deleterious mutations in the human genome. *PLoS Genet.* **17**, e1009676 (2021).
59. Hedrick, P. W. & Garcia-Dorado, A. Understanding inbreeding depression, purging, and genetic rescue. *Trends Ecol. Evol.* **31**, 940–952 (2016).
60. Bonnet, T. et al. Genetic variance in fitness indicates rapid contemporary adaptive evolution in wild animals. *Science* **376**, 1012–1016 (2022).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

Methods

Samples

Samples were collected from zoos, zoological museums, research institutes and farms from all over the world. Samples were provided from collaborators for research that was undertaken at the Natural History Museum of Denmark, permit 2020-12-7186-00733 from the Danish Ministry of Environment and Food, and when applicable, CITES Certificate of Scientific Exchange number DK003. Genomic DNA was extracted using DNeasy Blood and Tissue Kits (Qiagen) following the manufacturer's instructions. BGIseq libraries were built in China National GeneBank (CNGB), Shenzhen, China, and whole-genome paired-end sequencing (read length 2 × 100 bp) were performed on the BGISEQ500 platform. We aimed for 60–80× raw sequence coverage per sample. A total of 68 species for which a reference genome was available were retained in the final dataset, representing 151 trios for which whole blood or other tissue material was available for DNA extraction and for which parentage had been genetically determined[61]. Information on the samples is provided in Supplementary Table 1.

GMR estimation

We applied a similar bioinformatic analysis pipeline to our previous study of rhesus macaques[12]. Raw reads were trimmed with SOAPnuke filter[62]. The mapping was conducted with BWA-MEM version 0.7.15 (ref. [63]). The versions of the reference genomes for each species are provided in Supplementary Table 9. A post-mapping step removed any reads mapping to multiple regions of the genome as well as duplicated reads using Picard MarkDuplicates 2.7.1. We called variants for each individual using HaplotypeCaller in BP-RESOLUTION mode with GATK 4.0.7.0 (ref. [64]). This mode returns a genotype quality and depth for all positions of the genome, not only the polymorphic sites. As recommended by GATK best practices, GenomicsDBImport combined all gVCF files per species into a single file and GenotypeGVCF applied a joint genotyping of all samples within a given species (see Supplementary Table 3 with details of raw sequences coverage, mapping quality, and coverage after mapping and variant calling). Similar filtering methods to those in our previous study were then applied to detect DNMs[12]. Therefore, each trio was filtered as followed:

(1) For site filtering, the variant positions were filtered with the following parameters: QualByDepth (QD) < 2.0, FisherStrand (FS) > 20.0, RMSMappingQuality (MQ) < 40.0, MQRankSum < −2.0, MQRankSum > 4.0, ReadPosRankSum < −3.0, ReadPosRankSum > 3.0 and StrandOddsRatio (SOR) > 3.0 according to previously tested filters[12].

(2) For Mendelian violations, variants that deviated from Mendelian inheritance were selected using GATK SelectVariant and refined with an R script to keep only sites in which both parents were homozygous for the reference allele (HomRef), and the offspring was heterozygous (Het).

(3) For allelic balance filter, in the case of a DNM, approximately 50% of the reads in the offspring should support the alternative alleles. Our allelic balance filter cut-off was 30–70% of the reads supporting the alternative allele, similar to previous studies[12,32,65,66].

(4) For depth filter (DP), only positions with a DP > 0.5 × m_{depth} and DP < 2 × m_{depth} for each individual were kept, with m_{depth} being the average depth of the trio. This strict DP filter minimized the effects of sequencing errors in regions of low sequencing depth and mismapping errors in high-coverage regions.

(5) For genotype quality filter (GQ), to ensure that only high-quality genotypes were retained for the analysis of trios, we removed all sites where one individual of the trio had a GQ < 60 (see Supplementary Fig. 2 for a comparison of various GQ thresholds on a subset of species).

In addition, we called variants with bcftools (version 1.2)[67] in the region of the candidate DNMs and removed the sites that appeared as false-positive calls (that is, at least one parent had the same variant as the offspring or the offspring had no variant). The number of candidates discarded varied among species (Supplementary Table 2). This quality control step produced similar results to a manual check with IGV[68]. Moreover, calling variants with different variant callers has been shown to be an efficient method to reduce false-positive calls[2]. All positions of DNMs are provided in Supplementary Table 6. In addition, we showed that sample type, reference genome quality and mapping quality can affect the results on the number of candidates, the false-positive rate and false-negative rate (FNR), yet, the estimated mutation rates are not affected (Supplementary Figs. 3–5).

To estimate per-generation rates, we divided the number of candidate DNMs, without the apparent false-positive candidates, per the callable genome. A site was considered callable when it passed the same filters as the polymorphic sites, that is, when both parents were HomRef (filter 2) and the three individuals passed the depth filter (filter 4) and the genotype quality threshold (filter 5). On the sites considered callable, we applied a correction for the FNR, that is, the proportion of sites where true DNMs will not be called as such. Two methods have been used in the literature to estimate FNR: one is the simulation of mutations and the other is a correction on the filters that are not accounted for in the callable genome. As in our previous study of GMR[12], we used the latter method, which is more conservative. This corrected for the remaining filters that can only be applied on polymorphic sites, such as the site filters and the allelic balance filter (filter 2). We estimated the proportion of sites that would be filtered away by the site filters on the parameters following a known distribution (FS, MQRankSum and ReadPosRankSum), and the expected sites filtered away by the allelic balance filter as the number of true heterozygote sites (one parent HomRef, the other parent HomAlt and their offspring Het) outside the allelic balance threshold. The mutation rate per site per generation was then estimated per trio as $\mu_{\text{generation}}$ = DNMs/((1 − FNR) × 2 × CG). We estimated the 95% binomial confidence interval per species using the binconf() function in R, with the default Wilson score.

To calculate yearly rates (μ_{yearly}), we divided the per-generation rate by the average age of the parents at the time of reproduction weighted by the relative contribution of each parent (inferred with α for 105 trios) or by the generation time (for 46 trios without parental ages). The resulting μ_{yearly} estimates were averaged per species (for 29 species with multiple trios available). These yearly rates are dependent on the age of reproduction of the parents. Therefore, to calculate a yearly rate at generation time, we first modelled how the mutation rate of a trio was affected by the weighted average of the parental ages (using the paternal fraction estimated for that species as a weight). We then extended the model to fit how each species deviated from the average and used this to correct for differences between the observed reproductive age in our dataset and the expected generation time of a species (see Supplementary Note 1). With this, we estimated a new $\mu_{\text{yearly_modelled}}$ and a $\mu_{\text{generation_modelled}}$ that are more representative of the rate at generation time for each species.

Phylogenetic analysis

The phylogeny was built based on two sets of UCEs: 5,472 baits for 5,060 UCEs in tetrapods[57] and 2,628 baits for 1,314 UCEs in acanthomorphs[69]. We used the Phyluce software[70] to locate the probes in the reference genomes of our 68 species with 6 additional species contained in our original dataset. We extracted a flanking region of ±1,000 bp for each probe and aligned them with Mafft aligner version 7.470 (ref. [71]). We then created a 75% completion matrix, that is, each alignment contains at least 75% of the taxa (55 species), resulting in 63 alignments from the acanthomorph set and 2,742 probes from the tetrapod set (all alignments are available on Figshare). A phylogenetic tree was built using IQ-TREE version 2.0.3 (ref. [72]), with the appropriate substitution model inferred for each of the 2,805 alignments, a maximum likelihood tree search and 1,000 bootstrap replicates. To validate our tree, we also

estimated a second tree based on a MultiZ alignment to the human genome and obtained similar results (Extended Data Fig. 9). The phylogenetic tree was calibrated to absolute time using the chronos function of the 'ape' package in R, with a smoothing parameter lambda of 0 and a 'relaxed' model[73,74]. Fourteen nodes were calibrated following previously published calibrations. The robustness of the tree was assessed by removing each node independently (see Extended Data Fig. 3).

(1) Actinopterygii/Sarcopterygii: divergence time 416 million years ago (Ma), upper bound 425.4 Ma[75]
(2) The first node in the Actinopterygii group: divergence time 378.2 Ma[76]
(3) Sauropsida (birds and reptiles)/Synapsida (mammals): divergence time 313.4 Ma[77]
(4) Archosauria (birds)/Testudines: divergence time 260 Ma[78]
(5) The basal nodes of the Lepidosauria: divergence time 222.8 Ma[79]
(6) First mammalian node, Eutheria/Metatheria: divergence time 160.7 Ma[75]
(7) Galloanserae/Neoaves: divergence time 66 Ma[77]
(8) Glire/Primates: divergence time 61.7 Ma[77]
(9) Basal gekkotan node: divergence time 54 Ma[80]
(10) Passeriformes/Psittaciformes: divergence time 51.81 Ma[81]
(11) Cynoglossidae/Paralichthyidae: divergence time 50 Ma[76]
(12) *Sus scrofa*/other Cetartiodactyla: divergence time 48.5 Ma[77]
(13) Canidae/Arctoidea: divergence time 37.1 Ma[75]
(14) Hominoidea/Cercopithecoidea: divergence time 23.5 Ma[77]

Mutational spectrum and sex bias

To analyse the spectrum of mutation, we grouped the trios into higher taxonomic levels, that is, mammals, birds, fishes and reptiles. Thus, the percentages reported are based on the total candidate mutations from each group of species. We explored the genomic context of the mutations from a C or a G base to determine whether they were located in CpG sites (respectively followed by a G or preceded by a C) (see Supplementary Table 4). We phased the DNMs to their parental origin using the read-backed phasing method described previously (GitHub: https://github.com/besenbacher/POOHA)[82]. This method uses the read-pairs containing a DNM and another heterozygous variant to determine the parental origin of the mutation when the heterozygous variant is present in both the offspring and one of the parents. The phasing allowed us to identify parental biases in the contribution of the DNMs by grouping multiple species to increase the number of phased mutations and obtain a minimum of 30 phased mutations per taxon. From this analysis, we omitted the Egyptian roussette (*Rousettus aegyptiacus*), Chinese tree shrew (*Tupaia belangeri*), griffon vulture (*Gyps fulvus*), blue-throated macaw (*Ara glaucogularis*), snowy owl (*Bubo scandiacus*) and Darwin's rhea (*Rhea pennata*), as these could not be grouped with another monophyletic clade. To quantify the effect of parental age, a linear regression between the per-generation mutation rate and the average parental age at the time of reproduction was implemented using the lm function in R. Multiple linear regression was also used to identify whether paternal or maternal age was the strongest predictor of the empirical mutation rate.

Life-history trait analysis

We tested the effect of various life-history traits (fitted as continuous and discrete variables) on the yearly rate for each species using PGLS analysis in the R package 'caper'[83] (see Supplementary Table 9 for details about each life-history trait).

Effective population size

We used pairwise sequentially Markovian coalescent (PSMC) models to estimate the effective population size of each species[84]. Fastq sequences were obtained using bam format aligned sequences of one randomly selected father per species and were converted into fastq format using samtools mpileup command and vcf2fq. As recommended, the minimum depth was set to one-third of the average for the sample and twice the average for the maximum. For mammals, fish and reptiles, the parameters of the PSMC were set to −N25 for the maximum number of iterations of the algorithm, −t15 as the upper limit for the time to the most recent common ancestor, −r5 for the initial θ/ρ value, and finally the atomic intervals −p of '4 + 25 × 2 + 4 + 6'. These parameters were used previously for PSMC analysis of various species, including primates[84,85], cetaceans[86], Felidae[87], fishes[88,89] and turtles[90]. For birds, we used different parameters according to the literature with −N30 −t5 −r5 (ref. [91]). Finally, to simulate the history inferred by PSMC, we parameterized the generation time and the mutation rate inferred from the UCE alignment. We then explored the effect of the harmonic mean N_e over windows of 30,000 years to 1,000,000 years. We also compared N_e estimated obtained with this method with those estimated based on $N_e = \pi/4\mu$. Nucleotide diversity (π) was calculated using ANGSD[92]. This approach was implemented in three consecutive steps. From the alignment files, a global estimate of the site frequency spectrum was inferred using a maximum likelihood method, then the empirical π value was estimated per site, and finally, a sliding window approach was used to estimate π for each species. We used a window size of 50 kb and a step size of 10 kb together with an average pairwise estimation of the π to obtain global estimates of π. This analysis was restricted to unrelated individuals from each species, which corresponded to the 2 unrelated parents for 55 species, between 3 and 7 individuals for 10 species, and 3 species were excluded from this analysis as the parents were first-degree relatives.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Whole-genome sequences of all species except humans are accessible in the National Center for Biotechnology Information under the BioProject ID PRJNA767781. The human sequences are available on request to L.A.B. and should be used only for GMR studies, based on the participant's request. The alignments for the UCE tree are available on Figshare (https://doi.org/10.6084/m9.figshare.19221693.v1). All animal silhouettes are from PhyloPic (http://phylopic.org/), except for the silhouette of *S. scovelli*, which was created by J.S. The silhouette of *P. troglodytes* was created by T. M. Keesey (vectorization) and T. Hisgett (photography), and the one of *S. harrissi* silhouettes was created by S. Werning; both are available under a CC-BY 3.0 license (https://creativecommons.org/licenses/by/3.0/); the other silhouettes are available under a Public Domain Mark 1.0 licence.

Code availability

The bioinformatics pipeline to analyse the genomes and all other data analyses are available on GitHub (https://github.com/lucieabergeron/vertebrate_rate).

61. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
62. Chen, Y. et al. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience* **7**, 1–6 (2017).
63. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
64. Poplin, R. et al. Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at *bioRxiv* https://doi.org/10.1101/201178 (2018).
65. Kong, A. et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
66. Besenbacher, S. et al. Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat. Commun.* **6**, 5969 (2015).
67. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
68. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

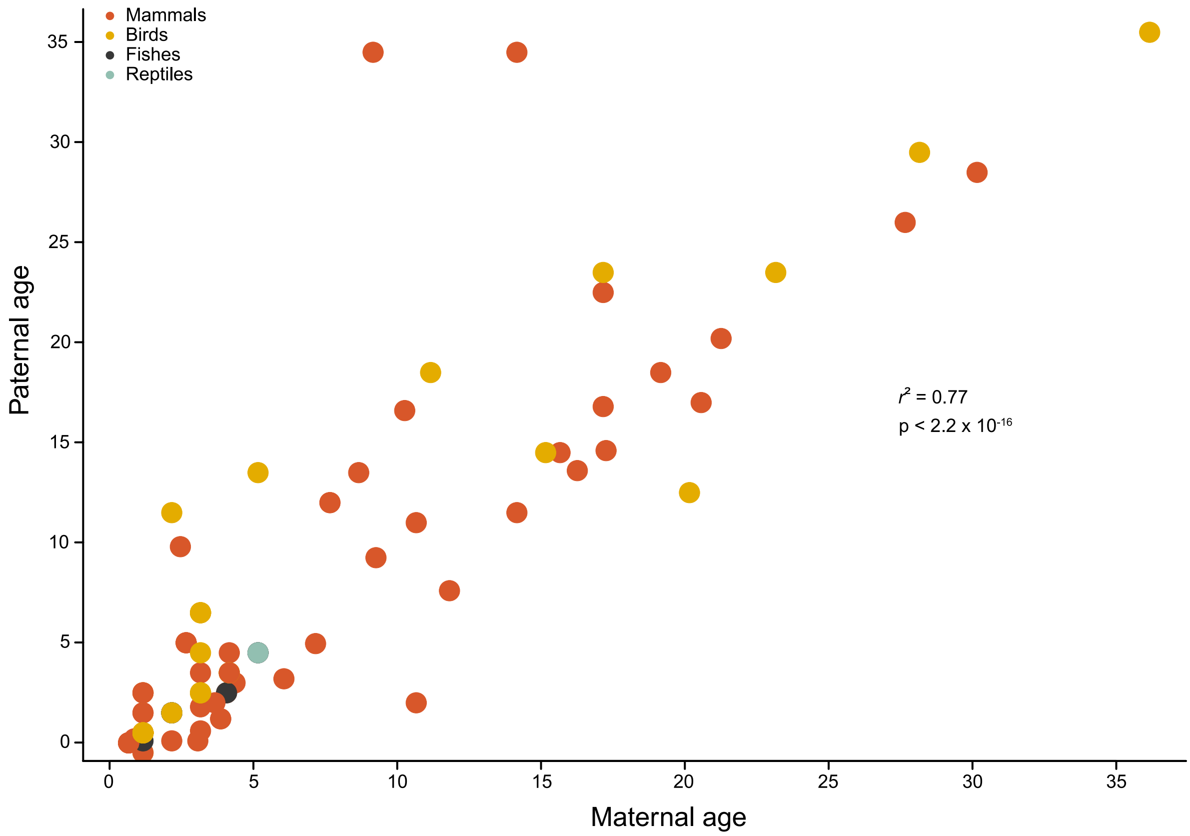
69. Alfaro, M. E. et al. Explosive diversification of marine fishes at the Cretaceous–Palaeogene boundary. *Nat. Ecol. Evol.* **2**, 688–696 (2018).
70. Faircloth, B. C. PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics* **32**, 786–788 (2016).
71. Katoh, K., Misawa, K., Kuma, K. I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
72. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
73. Sanderson, M. J. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol.* **19**, 101–109 (2002).
74. Kim, J. & Sanderson, M. J. Penalized likelihood phylogenetic inference: bridging the parsimony-likelihood gap. *Syst. Biol.* **57**, 665–674 (2008).
75. Meredith, R. W. et al. Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science* **334**, 521–524 (2011).
76. Hughes, L. C. et al. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. *Proc. Natl Acad. Sci. USA* **115**, 6249–6254 (2018).
77. Benton, M. J. & Donoghue, P. C. J. Paleontological evidence to date the tree of life. *Mol. Biol. Evol.* **24**, 26–53 (2007).
78. Green, R. E. et al. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. *Science* **346**, 1254449 (2014).
79. Sues, H. D. & Olsen, P. E. Triassic vertebrates of Gondwanan aspect from the Richmond basin of Virginia. *Science* **249**, 1020–1023 (1990).
80. Bauer, A. M., Böhme, W. & Weitschat, W. An Early Eocene gecko from Baltic amber and its implications for the evolution of gecko adhesion. *J. Zool.* **265**, 327–332 (2005).
81. Gelabert, P. et al. Evolutionary history, genomic adaptation to toxic diet, and extinction of the Carolina parakeet. *Curr. Biol.* **30**, 108–114.e5 (2020).
82. Maretty, L. et al. Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature* **548**, 87–91 (2017).
83. Orme, D. et al. The caper package: comparative analysis of phylogenetics and evolution in R. R version 1.0.1 https://cran.r-project.org/package=caper (2018).
84. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
85. Schmitz, J. et al. Genome sequence of the basal haplorrhine primate *Tarsius syrichta* reveals unusual insertions. *Nat. Commun.* **7**, 12997 (2016).
86. Vijay, N. et al. Population genomic analysis reveals contrasting demographic changes of two closely related dolphin species in the last glacial. *Mol. Biol. Evol.* **35**, 2026–2033 (2018).
87. Liu, Y. C. et al. Genome-wide evolutionary analysis of natural history and adaptation in the world's tigers. *Curr. Biol.* **28**, 3840–3849.e6 (2018).
88. Xu, S., Zhao, L., Xiao, S. & Gao, T. Whole genome resequencing data for three rockfish species of Sebastes. *Sci. Data* **6**, 97 (2019).
89. Yuan, Z. et al. Historical demography of common carp estimated from individuals collected from various parts of the world using the pairwise sequentially markovian coalescent approach. *Genetica* **146**, 235–241 (2018).
90. Fitak, R. R. & Johnsen, S. Green sea turtle (*Chelonia mydas*) population history indicates important demographic changes near the mid-Pleistocene transition. *Mar. Biol.* **165**, 110 (2018).
91. Nadachowska-Brzyska, K., Li, C., Smeds, L., Zhang, G. & Ellegren, H. Temporal dynamics of avian populations during pleistocene revealed by whole-genome sequences. *Curr. Biol.* **25**, 1375–1380 (2015).
92. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).
93. Milholland, B. et al. Differences between germline and somatic mutation rates in humans and mice. *Nat. Commun.* **8**, 15183 (2017).
94. The 1000 Genomes Project. Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* **43**, 712–714 (2011).
95. Rahbari, R. et al. Timing rates and spectra of human germline mutation. *Nat. Genet.* **48**, 126–133 (2015).
96. Wong, W. S. W. et al. New observations on maternal age effect on germline de novo mutations. *Nat. Commun.* **7**, 10486 (2016).
97. Turner, T. N. et al. Genomic patterns of de novo mutation in simplex autism. *Cell* **171**, 710–722.e12 (2017).
98. Sasani, T. A. et al. Large three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *eLife* **8**, e46922 (2019).
99. Kessler, M. D. et al. De novo mutations across 1465 diverse genomes reveal mutational insights and reductions in the Amish founder population. *Proc. Natl Acad. Sci. USA* **117**, 2560–2569 (2020).
100. Malinsky, M. et al. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nat. Ecol. Evol.* **2**, 1940–1955 (2018).
101. Koch, E. M. et al. De novo mutation rate estimation in wolves of known pedigree. *Mol. Biol. Evol.* **36**, 2536–2547 (2019).
102. Harland, C. et al. Frequency of mosaicism points towards mutation-prone early cleavage cell divisions in cattle. Preprint at *bioRxiv* https://doi.org/10.1101/079863 (2017).
103. Pfeifer, S. P. Direct estimate of the spontaneous germ line mutation rate in African green monkeys. *Evolution* **71**, 2858–2870 (2017).

Acknowledgements The authors thank the following contributors of samples for this study: A. Girard, C. Small, E. Couture, E. Gangloff, A. Bronikowski, F. Yu, H. Fernández, A. Carbajal Brossa, the Barcelona Zoo Biological Bank, J. Partecke, J. Judson, F. Janzen, J. Fjeldså, K. Thorup, K. Glover, L. Koren, M. Nagel, M. Fredholm, M. Liedvogel, T. Aquarium, P. Vullioud, S.-J. Luo, T. Gamble, Y. Yovel, J. Bakker, C. Bombis, T. Charlton, A. Corl, A. Foote, N. Geli, M. Guille, K. L. Hansen, W. Huizinga, M. Hunter, T. Knauf-Witzens, T. Lund Koch, S. Potier, A. Prahl, K. Robertson, C. Scala, M. Schellerup, I. Schnell, K. Vesterdorf, K. Wendelin, K. Worm and W.-z. Wang; G. Pacheco for valuable advice in the laboratory; K. Boomsma for stimulating conceptual discussions and for providing comments on the final version of this manuscript; and GenomeDK at Aarhus University for providing computational resources and support for this study. All sequencing data were generated with MGI-sequencers at the China National Genebank of BGI-Shenzhen. This project was funded by the Strategic Priority Research Programme (XDB13020000) and the International Partnership Programme (no. 152453KYSB20170002) of the Chinese Academy of Sciences, a Villum Investigator grant (no. 25900) from The Villum Foundation, and a Carlsberg Foundation Grant to G.Z. (CF16-0663). L.A.B. was supported by the Carlsberg Foundation and the Villum Foundation. M.H.S. was funded by the Novo Nordisk Foundation (NNF18OC0031004). J.I.H. was funded by the German Research Foundation (DFG) as part of the SFB TRR 212 (NC³) (project nos. 316099922 and 396774617), the priority programme "Antarctic Research with Comparative Investigations in Arctic Ice Areas" SPP 1158 (project no. 424119118) and the sequencing costs in projects scheme (project no. 497640428).

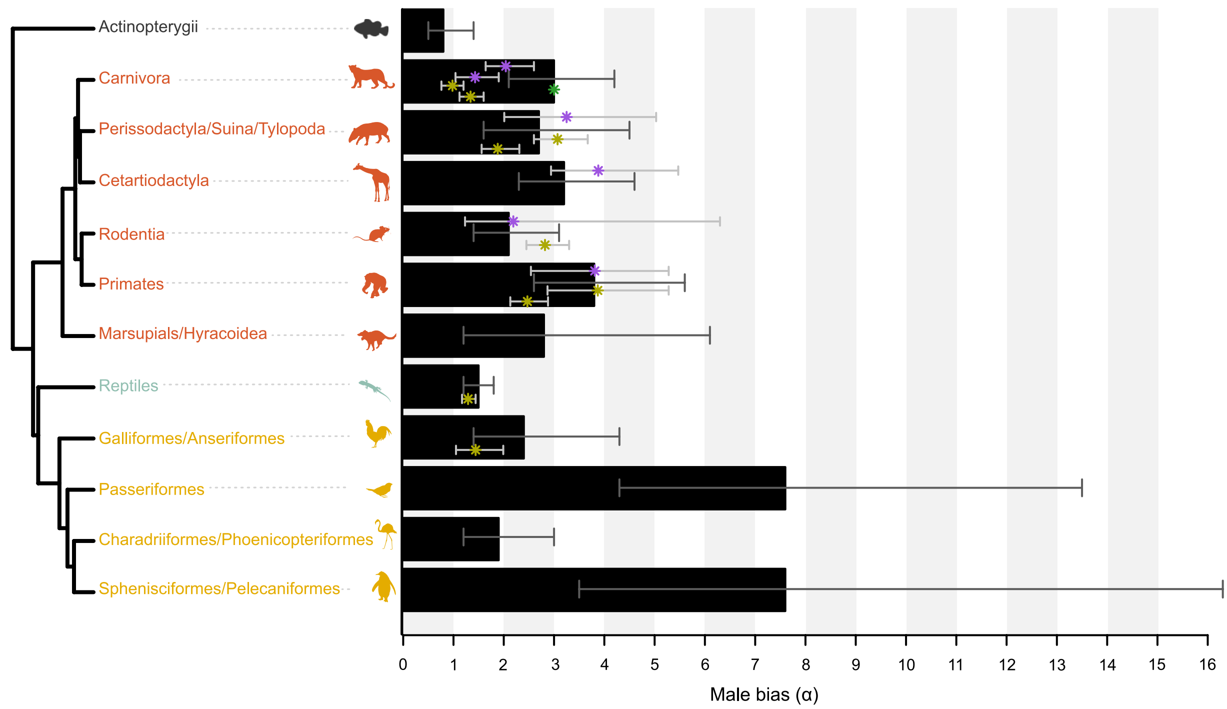
Author contributions G.Z., M.H.S., S.B. and L.A.B. conceived this work. M.F.B., B.Q., J.I.H., Z.L., J.S.L. and C.S. provided samples for several species, as well as input into the writing and results interpretation. L.A.B., J.Z., P.L. and M.T.P.G. participated in the extraction, library preparation and sequencing. All analyses were conducted by L.A.B. with input from J.S. for the phylogenetic analysis and S.B. for the mutation rate estimation. L.A.B., G.Z., S.B., M.H.S. and J.I.H. wrote the initial draft of the manuscript with input from all co-authors. G.Z. supervised this project.

Competing interests The authors declare no competing interests.

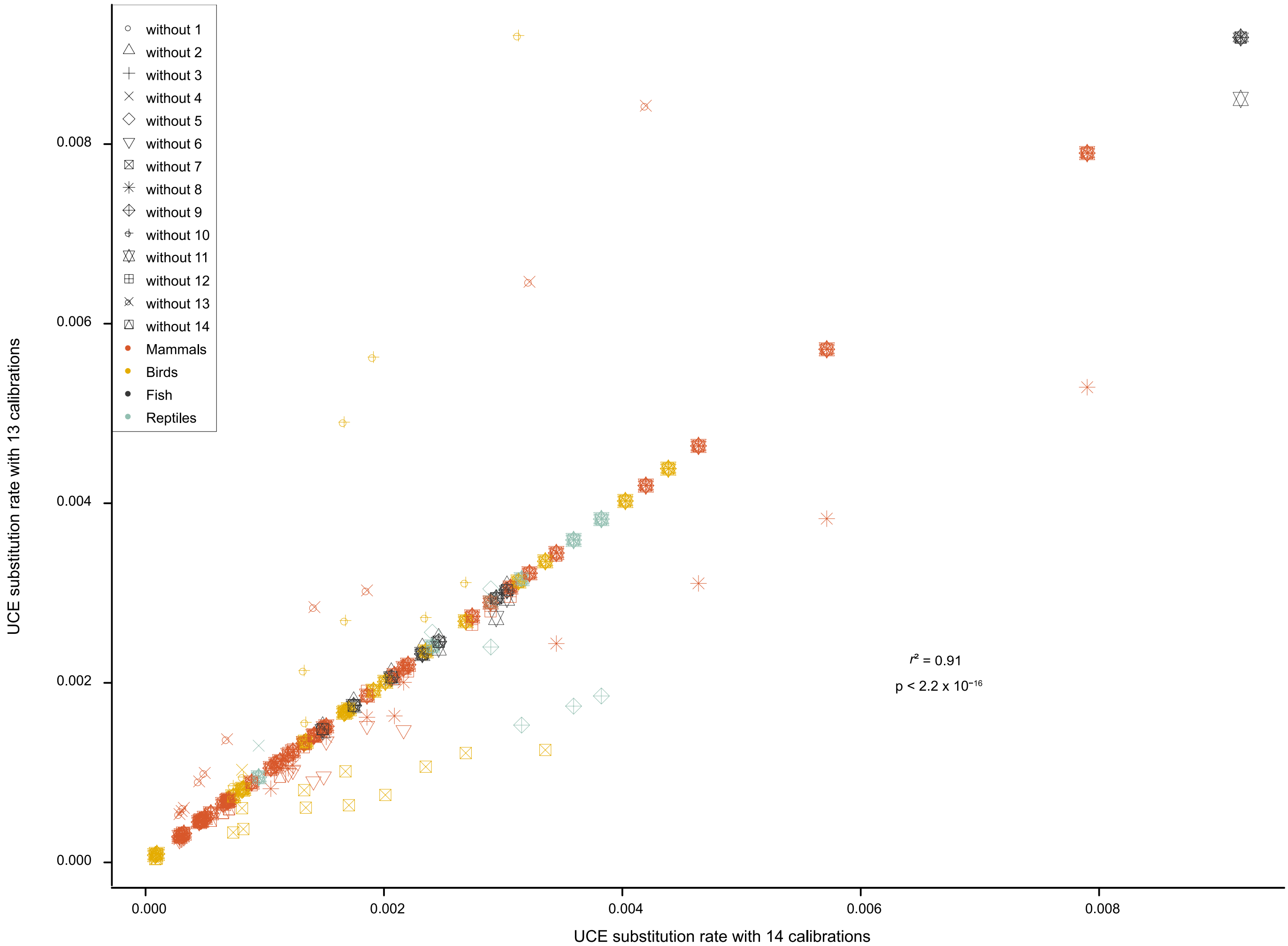
Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-05752-y.
Correspondence and requests for materials should be addressed to Lucie A. Bergeron or Guojie Zhang.
Peer review information *Nature* thanks Anne Goriely and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.
Reprints and permissions information is available at http://www.nature.com/reprints.



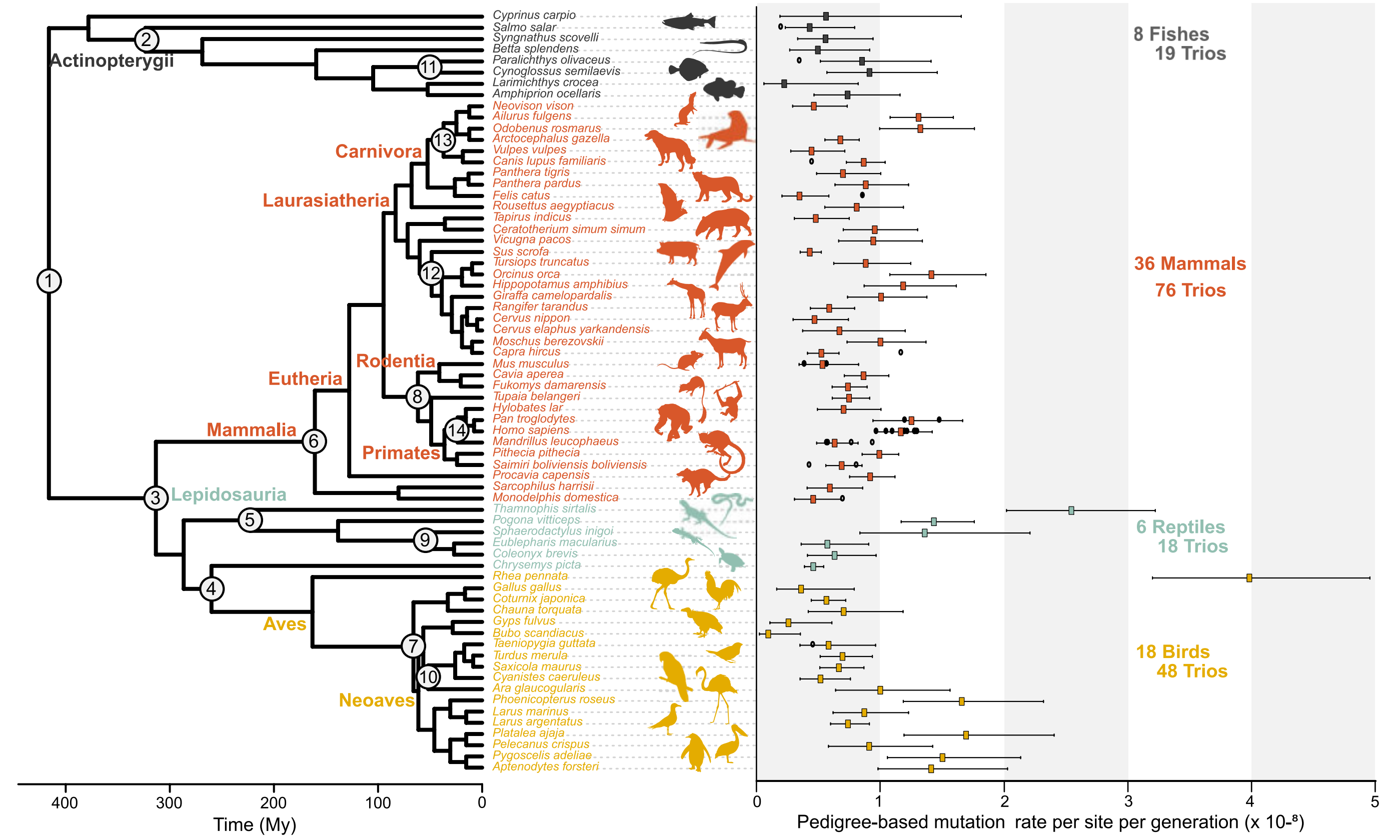
Extended Data Fig. 1 | Association of parental ages. Maternal and paternal ages are significantly positively correlated for the 105 trios with known parental age at reproduction (linear regression; adjusted $r^2 = 0.77$, $F = 342.3$ on 1 and 103 DF, $p < 2.2 \times 10^{-16}$).



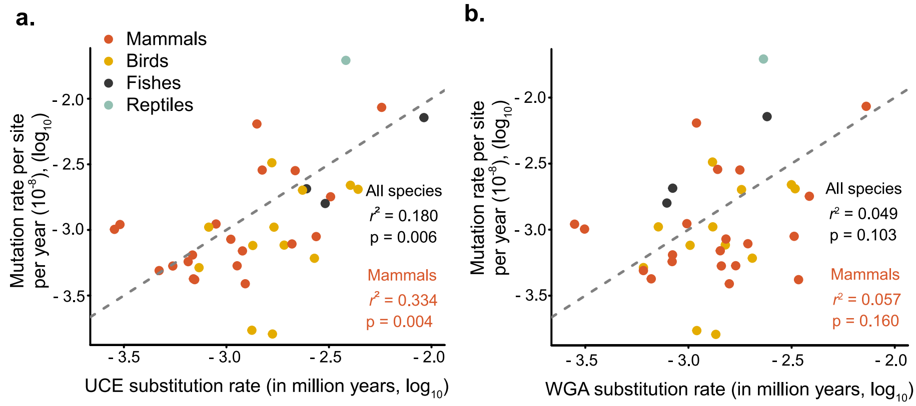
Extended Data Fig. 2 | Comparison of published male bias estimates (α) using genome alignments and our male bias estimates (modified Fig. 1c of the main text). The yellow points are α estimates from Wilson Sayres et al.[28], and the purple points are α estimates from Wu et al.[31]. Most of the common species reveal similar estimates with overlapping 95% confidence intervals. However, the estimates of α based on genome alignments are generally lower for dogs and cats than our estimates, yet the pedigree-based estimate of α for cats (Wang et al.[20]; green point) is similar to our estimate. See also Supplementary Table 5. The barplots represent male biases estimated by clustering different species per group (to have a minimum of 30 phased mutations per group) and the 95% confidence intervals were based on the binomial distribution. The silhouette of *Sygnathus* was created by J.S. All other silhouettes are from PhyloPic (http://phylopic.org), except one of the silhouettes of *Sarcophilus harrissi*, which was created by S. Werning, and the silhouette of *Pan troglodytes*, which was created by T. M. Keesey (vectorization) and T. Hisgett (photography); both are available under a CC-BY 3.0 licence (https://creativecommons.org/licenses/by/3.0).



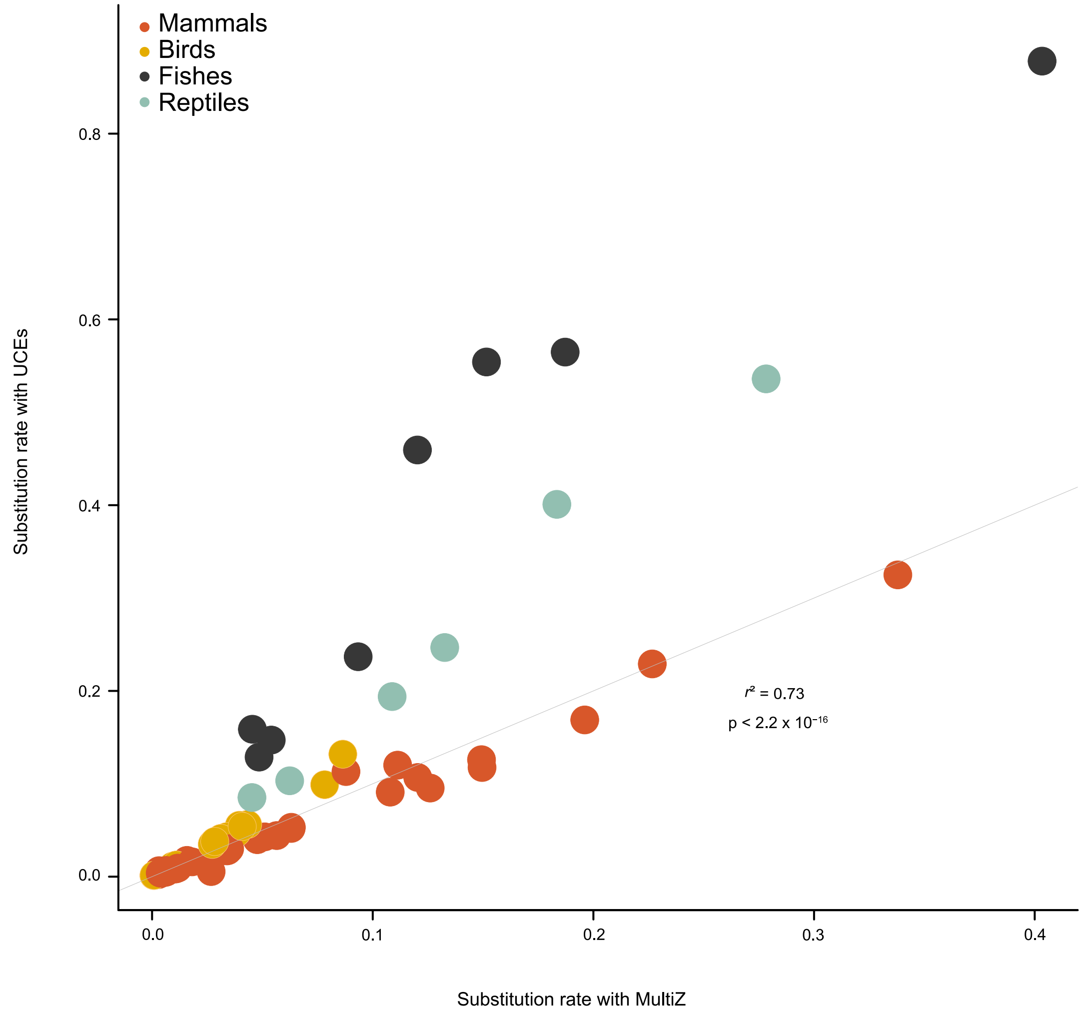
Extended Data Fig. 3 | Robustness of the calibration. We compared the estimated substitution rates using the 14 initial calibration points with the inferred substitution rates using only 13 calibration points (with 14 iterations to remove each calibration node one by one). We found a strong relationship between the rates estimated with 14 and 13 calibrations (linear regression adjusted r^2 = 0.91, F = 9416 on 1 and 950 DF, p-value: < 2.2×10^{-16}). However, some of the calibration points had a stronger impact on the estimated substitution rates. For instance, removing the two bird nodes (7 and 10), the gekko node (9), the Canidae/Arctoidea node (13) and the Glires/Primate node (8) altered some of the substitution rate estimates.



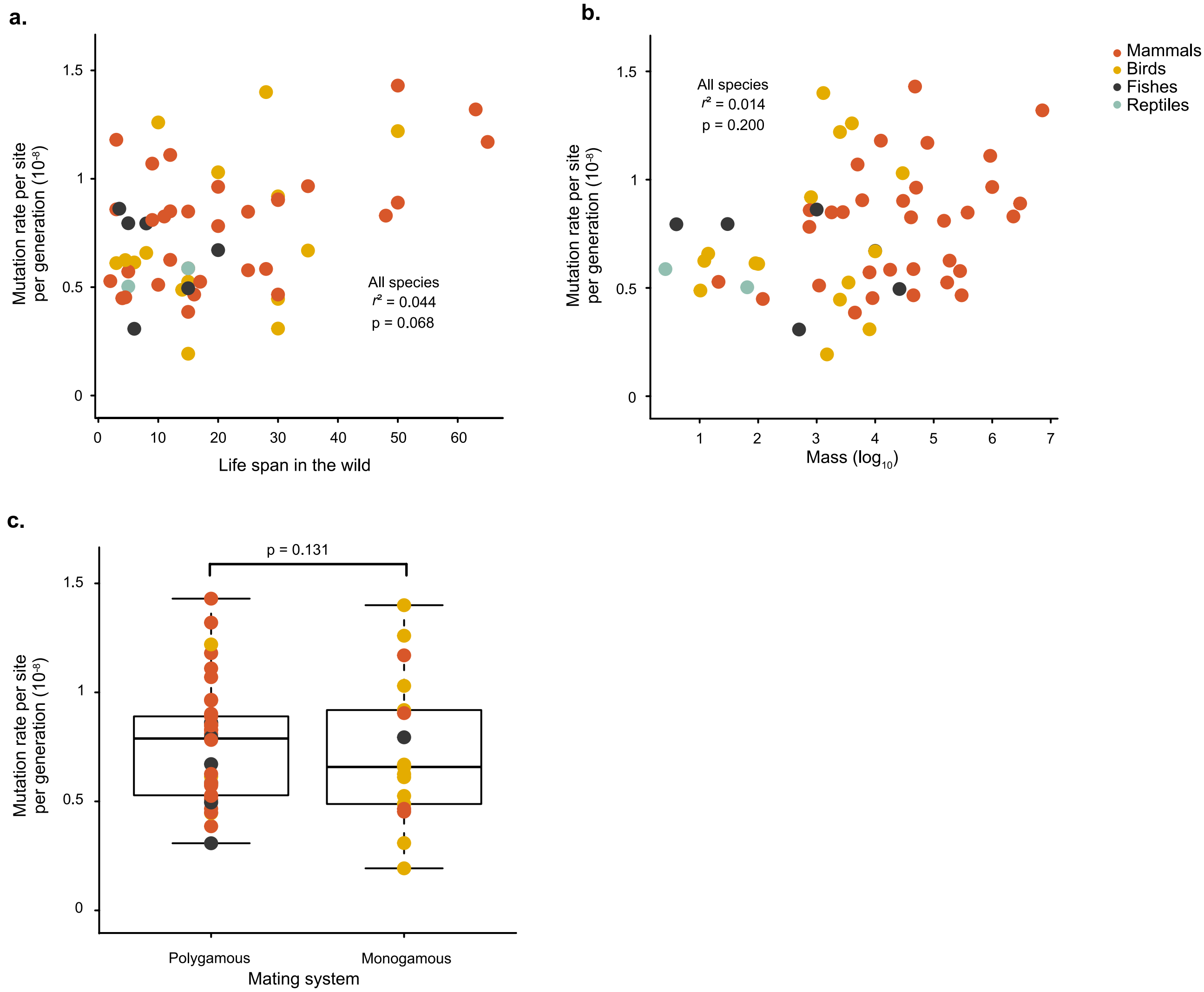
Extended Data Fig. 4 | Per-generation mutation rates (similar to Fig. 1a) including published data on closely related species. For each species, the colored squares represent the average per-generation observed rate, along with the 95% confidence intervals based on the binomial distribution, and the black points represent published estimates from similar or closely related species to those included in our dataset. For most of the species, these estimates lie within the 95% confidence intervals of our estimates. Published estimates are from: *Felis catus* (Wang et. al.[20]), *Mus musculus* (Milholland et al.[93], Lindsay et al.[34]), *Pan troglodytes* (Venn et al.[21], Tatsumoto et al.[23], Besenbacher et al.[16]), *Homo sapiens* (Conrad et al.[94], Kong et al.[65], Francioli et al.[32], Rahbari et al.[95], Wong et al.[96], Jónsson et al.[22], Maretty et al.[82], Turner et al.[97], Sasani et al.[98], Kessler et al.[99]). The closely related species are from: close to the *Salmo salar*, *Clupea harengus* (Feng et al.[51]), close to *Paralichthys olivaceus*, the Cichlid (Malinsky et al.[100]), close to *Canis lupus familiaris*, *Canis lupus* (Koch et al.[101]), close to *Capra hircus*, *Bos taurus* (Harland et al.[102]), close to *Mandrillus leucophaeus*, *Papio anubis* (Wu et al.[13]), *Macaca mulatta* (Wang et al.[14], Bergeron et al.[12]), and *Chlorocebus sabaeus* (Pfeifer[103]), close to *Saimiri boliviensis boliviensis*, *Aotus nancymaae* (Thomas et al.[17]), close to *Monodelphis domestica*, *Ornithorhynchus anatinus* (Martin et al.[49]), close to *Taeniopygia guttata*, *Ficedula albicollis* (Smeds et al.[50]). See also Supplementary Table 8. The silhouette of *Sygnathus* was created by J.S. All other silhouettes are from PhyloPic (http://phylopic.org), except one of the silhouettes of *S. harrissi*, which was created by S. Werning, and the silhouette of *P. troglodytes*, which was created by T. M. Keesey (vectorization) and T. Hisgett (photography); both are available under a CC-BY 3.0 licence (https://creativecommons.org/licenses/by/3.0).



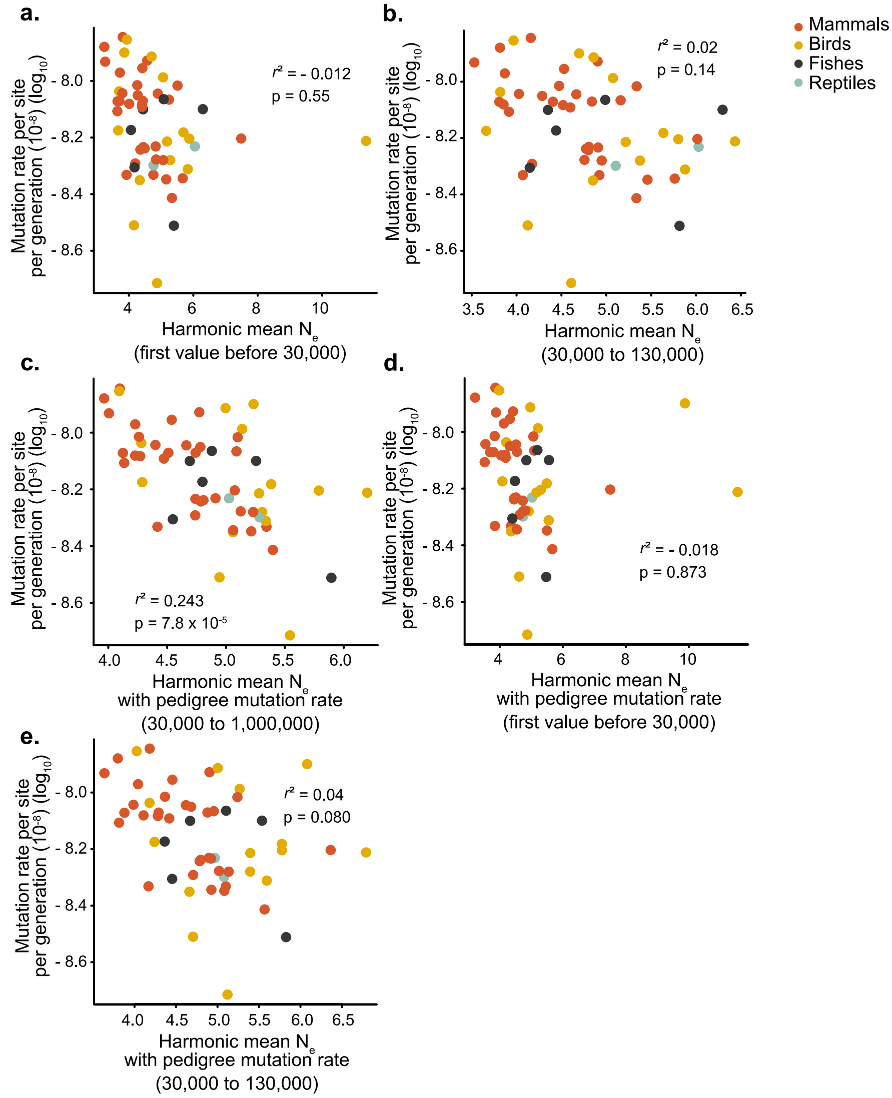
Extended Data Fig. 5 | Germline mutation rates are associated with long-term substitution rates. This figure is similar to the main Fig. 2 but uses phylogenetic regression (PGLS) on a log scale. The grey dashed lines indicate equality. **a**. Using a log scale, there is a significant positive correlation between the per-year rates and the rates derived from Ultraconserved elements (UCEs) and their flanking sequences. **b**. However, this correlation is not significant when comparing the per-year rates with the rates derived from the whole genome alignments (WGAs).



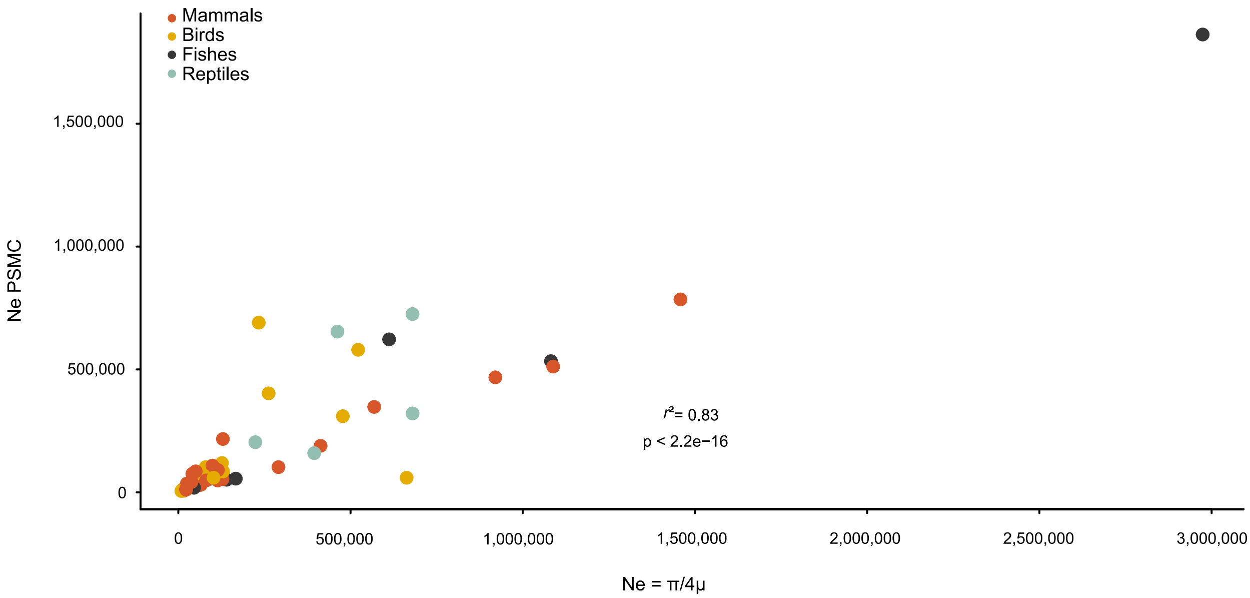
Extended Data Fig. 6 | Comparison of substitution rates estimated with Ultra Conserved Elements (UCEs) and MultiZ alignments. The substitution rates estimated with the two methods are highly correlated (linear regression: adjusted $r^2 = 0.73$, $F = 179.9$ on 1 and 66 DF, $p < 2.2 \times 10^{-16}$).



Extended Data Fig. 7 | Three life-history traits are not significantly associated with the per-generation mutation rate. a. lifespan in the wild, **b.** body mass and **c.** the mating system (polygamy versus monogamy). The total number of species with modeled per-generation rate was 55 for the phylogenetic regression (PGLS). The boxplots represent the median, the interquartile range, and the maximum and minimum excluding outliers.



Extended Data Fig. 8 | The drift barrier hypothesis on different times and different mutation rate parameters used to estimate N_e with phylogenetic regression (PGLS). **a**. The correlation between N_e and the mutation rate per generation is not significant when using the most recent value before 30,000 years estimated by PSMC. **b**. The relationship is also not significant when using the harmonic mean over a more recent period of time (30,000 years to 130,000 years ago). However, this relationship is significant for mammals (adjusted $r^2 = 0.104$, $p = 0.04$). We used the harmonic mean over the past million years in the main text, as PSMC is not reliable over recent periods. **c**. When looking at the relationship between the mutation rate and N_e, estimated using the pedigree-based mutation rate, we find a stronger signal over the past 1,000,000 years, probably due to the circularity of this analysis. **d**. However, the relationship is still not significant when using the most recent time point or **e**. the average over the past 100,000 years.



Extended Data Fig. 9 | Effective population sizes calculated with two different methods (see main text) are significantly correlated. The harmonic mean of the population size estimated with PSMC from 30,000 to 1,000,000 years ago is significantly correlated with the effective population size estimated from $N_e = \pi/4\mu$ (linear regression: adjusted $r^2 = 0.83$, $F = 316.3$ on 1 and 63 DF, $p < 2.2 \times 10^{-16}$).

nature portfolio

Corresponding author(s): Lucie A. Begeron and Guojie Zhang

Last updated by author(s): Jan 2, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection
Data analysis	The software used were: samtools 1.2 and 0.1.18, bcftools 1.2 and 1.9, bwa 0.7.15, Picard MarkDuplicates 2.7.1, R 3.5.1 and 4.0.2, IQTREE 2.0.3, Phyluce, ANGSD 0.920, soapnuke 1.5.6, python 3.7.3, java 1.8.0_222, gatk 4.0.7.0. The entire pipeline used is available at https://github.com/lucieabergeron/vertebrate_rate.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Whole-genome sequences of all species except humans are accessible in NCBI (National Center for Biotechnology Information) under the BioProject: PRJNA767781

(https://www.ncbi.nlm.nih.gov/bioproject/767781). The human sequences are available upon request. The alignments for the UCE tree are available on Figshare (https://doi.org/10.6084/m9.figshare.19221693.v1).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Not applicable
Population characteristics	Not applicable
Recruitment	Not applicable
Ethics oversight	Not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study estimates the germline mutation rates for 68 species of vertebrates using whole genome comparison of pedigrees.
Research sample	We analyzed 151 trios for 68 species of vetebrates, including birds, reptiles, fishes and mammals.
Sampling strategy	We collected trios (mother, father, offspring) for species with reference genome available. Our sample size for each species was limited by the pedigree available.
Data collection	Lucie A. Bergeron collected all the data, with the help of collaborators from zoo, research centers and farm.
Timing and spatial scale	Data were collected between October 2017 and December 2018.
Data exclusions	After data analysis, some of the samples were not as related as stated by the samples providers, therefore, they were excluded.
Reproducibility	All sequences and scripts are provided for reproducibility of our results.
Randomization	This is not relevant to our study, randomization was not possible as the sample size was small per species.
Blinding	Blinding was not relevant to our study as there were no expected data most of the rates were never estimated before from pedigrees.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Article

Cardiogenic control of affective behavioural state

https://doi.org/10.1038/s41586-023-05748-8

Received: 31 December 2021

Accepted: 20 January 2023

Published online: 1 March 2023

Open access

Check for updates

Brian Hsueh[1,6], Ritchie Chen[1,6], YoungJu Jo[1], Daniel Tang[1], Misha Raffiee[1], Yoon Seok Kim[1], Masatoshi Inoue[1], Sawyer Randles[1], Charu Ramakrishnan[1], Sneha Patel[1], Doo Kyung Kim[1], Tony X. Liu[1], Soo Hyun Kim[1], Longzhi Tan[1], Leili Mortazavi[1], Arjay Cordero[1], Jenny Shi[1], Mingming Zhao[2], Theodore T. Ho[1], Ailey Crow[1], Ai-Chi Wang Yoo[1], Cephra Raja[1], Kathryn Evans[1], Daniel Bernstein[2], Michael Zeineh[3], Maged Goubran[3] & Karl Deisseroth[1,4,5] ✉

Emotional states influence bodily physiology, as exemplified in the top-down process by which anxiety causes faster beating of the heart[1–3]. However, whether an increased heart rate might itself induce anxiety or fear responses is unclear[3–8]. Physiological theories of emotion, proposed over a century ago, have considered that in general, there could be an important and even dominant flow of information from the body to the brain[9]. Here, to formally test this idea, we developed a noninvasive optogenetic pacemaker for precise, cell-type-specific control of cardiac rhythms of up to 900 beats per minute in freely moving mice, enabled by a wearable micro-LED harness and the systemic viral delivery of a potent pump-like channelrhodopsin. We found that optically evoked tachycardia potently enhanced anxiety-like behaviour, but crucially only in risky contexts, indicating that both central (brain) and peripheral (body) processes may be involved in the development of emotional states. To identify potential mechanisms, we used whole-brain activity screening and electrophysiology to find brain regions that were activated by imposed cardiac rhythms. We identified the posterior insular cortex as a potential mediator of bottom-up cardiac interoceptive processing, and found that optogenetic inhibition of this brain region attenuated the anxiety-like behaviour that was induced by optical cardiac pacing. Together, these findings reveal that cells of both the body and the brain must be considered together to understand the origins of emotional or affective states. More broadly, our results define a generalizable approach for noninvasive, temporally precise functional investigations of joint organism-wide interactions among targeted cells during behaviour.

Interoceptive processing of visceral physiological signals, such as cardiac palpitations or stomach fullness, is crucial for maintaining homeostasis[1–3]. Diverse psychiatric conditions, such as anxiety disorders, panic disorder, body dysmorphic disorders and addiction, have been hypothesized to be related to dysregulation of interoceptive monitoring by the brain[3,4], and can be statistically correlated with specific visceral organ dysfunction. For example, patients with panic disorder and agoraphobia are more likely to have mitral valve prolapse or clinical symptoms similar to paroxysmal supraventricular tachycardia[5,6]. Modern correlative studies have further suggested links between cardiac changes and affect regulation[7,8], including correlations between cardiac interoception with anxiety and functional alterations in the insular cortex, a cortical region that has a central role in both the processing of physiological signals and the regulation of emotions[4,10]. However, determining whether primary physiological signals such as increased heart rate can causally influence behavioural states, as proposed in classical physiological theories of emotion[9], has—although widely debated—remained largely experimentally intractable[11]. Available nonspecific interventions that might disrupt cardiac signals (such as electrical vagus nerve stimulation) are well known to also induce numerous physiological changes that would be unwanted in this context, including direct suppression of respiratory and heart rates as well as anxiolytic and antidepressive effects[11–13], giving rise to multiple direct confounds for the question explored here. Other nonspecific interventions to alter cardiac rhythms, such as broadly active pharmacological stimulants or electrical pacemakers[14], also introduce insuperable confounds through initial actions beyond the direct pacing of cardiomyocytes, and thus lack the necessary precision. Studying the key question of how cardiac physiology regulates emotional states has remained inaccessible, and the effects on behaviour remain unknown.

Precise modulation of electrochemical signals in the heart and other peripheral organs in vivo would enable fundamental studies of physiology and interoceptive signalling[15–19], but stimulation methods that operate with high spatial and temporal precision in highly dynamic

[1]Department of Bioengineering, Stanford University, Stanford, CA, USA. [2]Department of Pediatrics, Stanford University, Stanford, CA, USA. [3]Department of Radiology, Stanford University, Stanford, CA, USA. [4]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA. [5]Howard Hughes Medical Institute, Stanford University, Stanford, CA, USA. [6]These authors contributed equally: Brian Hsueh, Ritchie Chen. ✉e-mail: deissero@stanford.edu

environments such as the beating heart[20–28] are limited. Electrical pacemakers require invasive surgical implantation to deliver local indiscriminate stimulation that lacks cell-type specificity[20–23]. Optogenetics might in principle facilitate cardiomyocyte-specific control with high spatial and temporal precision[14], but existing optogenetic methods have been limited to acute demonstrations that require exposure or even excision of the heart to deliver light[23–28], all of which are incompatible with freely moving studies of behaviour. Thus far, to our knowledge, no study beyond the brain has achieved precise and noninvasive organ-level control of behavioural or physiological function. Establishing noninvasive approaches to manipulate physiology with cell-type specificity would enable long-sought functional studies of signals that arise from cells across the organism, and reveal the causal influences of these cells on brain function and behaviour.

Conventional microbial opsins have not been sensitive enough to control a large organ such as the heart with the requisite power to facilitate behavioural studies within intact animals[23–28], but the discovery of the highly sensitive and red-shifted pump-like channelrhodopsin ChRmine led us to consider the potential for noninvasive optogenetic control of deep tissue with minimal irradiance[29]. We previously found that ChRmine enabled neuromodulation of deep brain circuits without intracranial surgery[30], which raised the possibility that this optogenetic tool might be broadly applicable to modulating biological processes across the entire body of large organisms such as mammals. Specifically, we hypothesized that ChRmine might allow on-demand deep-tissue control of cardiac pacing when targeted to cardiomyocytes without the need for direct exposure of the heart[29,30].

Noninvasive optogenetic cardiac control

We first achieved cardiomyocyte-restricted expression by placing the ChRmine transgene under the control of the mouse cardiac troponin T promoter (mTNT), using the AAV9 serotype, which exhibits tropism for cardiac tissue[25,31]. Infection of cultured primary cardiomyocytes with AAV9-mTNT::ChRmine-2A-oScarlet enabled light-evoked contractions with irradiance as low as 0.1 mW mm^{-2}, consistent with the photosensitivity of ChRmine in neurons[29] (Extended Data Fig. 1 and Supplementary Video 1).

We next determined whether systemic viral gene delivery of ChRmine, despite the lower multiplicity of infection compared with transduction by direct local injection[30,32], could allow noninvasive control of heart rhythms in wild-type mice. Retro-orbital injection of AAV9 enabled restricted expression of ChRmine in cardiomyocytes throughout the heart, with homogeneous expression in both ventricular and atrial walls and no off-target expression in other cardiac cell types (fibroblasts and neuronal ganglia) or in other organs (Fig. 1a,b and Extended Data Fig. 2). When pulsed 589-nm light was delivered through intact skin overlying the thorax of anaesthetized mice, we observed robust photoactivation of cardiac QRS complexes within a safe range of irradiance comparable to that used for transcranial optogenetics[30] (Fig. 1c,d). Reliable cardiac rhythms were induced at even supraphysiological rates of up to 900 beats per minute (bpm), within the photophysical properties of ChRmine[33], with an immediate return to naturally paced sinus rhythm upon the cessation of light (Fig. 1e–g). This approach was not able to decrease heart rate below baseline levels, but afforded spatial control of cardiac rhythms by evoking either right or left ventricular pacing depending on the placement of the laser (Extended Data Fig. 3a–i).

To translate this approach to mouse behaviour, we mounted a 591-nm micro-LED onto a wearable fabric vest to deliver light through intact skin overlying the chest wall (Fig. 2a and Extended Data Fig. 4a,b). This integration of a molecular tool with accessible electronics enabled the initial demonstration of noninvasive and sustained ventricular pacing at experimenter-defined rhythms suitable for most behavioural assays in freely moving mice (Extended Data Fig. 4).

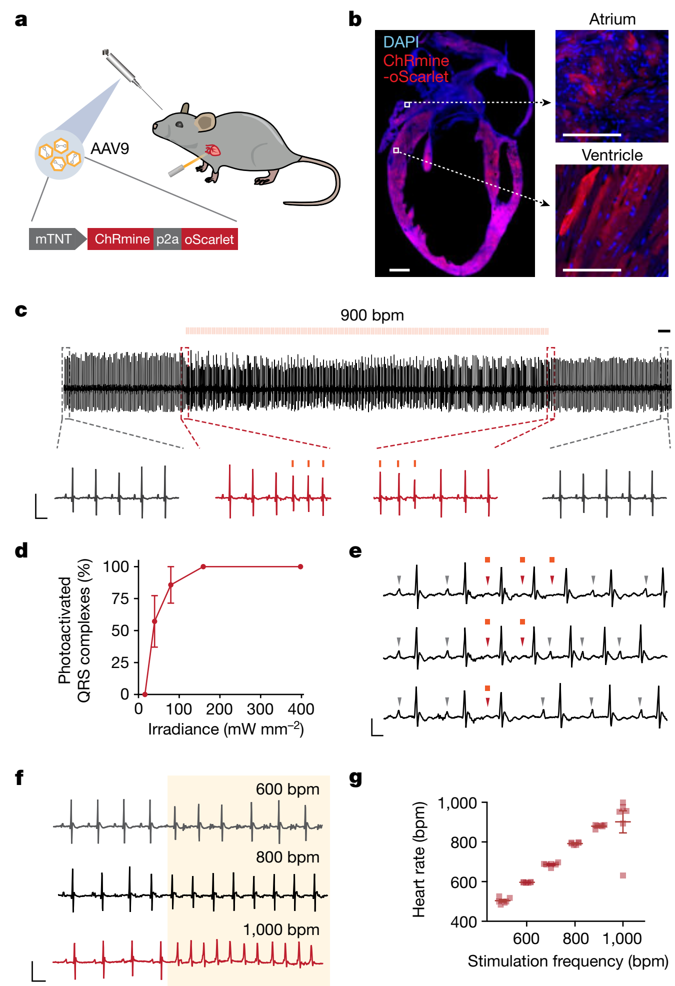


Fig. 1 | Development of a noninvasive optical pacemaker. a, Schematic showing the optical control of cardiac rhythm with an external light source enabled by retro-orbital injection of AAV9-mTNT::ChRmine-p2A-oScarlet. **b**, Confocal cross-section images indicating homogeneous transgene expression of ChRmine-p2A-oScarlet (red) with DAPI staining (blue) in atria and ventricles. Scale bars, 1 mm (main); 100 μm (inset). **c**, Example electrocardiogram (ECG) trace with optical pacing using 589 nm light delivered at 15 Hz (at 900 bpm) with a pulse width of 10 ms and irradiance of 160 mW mm^{-2}. Scale bar, 500 ms. Inset traces: ECG signal before and after light delivery (grey) and at light onset and cessation (red). Scale bar, 50 ms, 0.5 mV. **d**, Reliability of photoactivated QRS complexes at 900 bpm as a function of cutaneous optical irradiance ($n = 6$ mice). **e**, Example ECG traces of individual 10-ms optical pulses. Grey arrowheads indicate P waves associated with sinus rhythm, which are overridden (red arrowheads) during optical pacing. Scale bar, 25 ms, 0.25 mV. **f**, Example ECG traces of pacing at 600, 800 and 1,000 bpm. Scale bar, 50 ms, 0.5 mV. The shaded area indicates the period of illumination at the specified frequency at 100% duty cycle. **g**, Characterization of optical pacing fidelity, showing stimulation frequency versus ECG-measured heart rate ($n = 6$ mice). All ECG measurements were performed in anaesthetized mice. Data are mean ± s.e.m.

To test whether heart rhythms directly set by this optical pacemaker could influence behaviour, we optogenetically induced intermittent ventricular tachycardia (900 bpm for 500 ms every 1,500 ms) to mimic non-sustained arrhythmias that are observed during stressful contexts[34–36], while shortening the duration of decreases in systolic blood pressure and avoiding incidental heating from light-delivery devices (Fig. 2b and Extended Data Figs. 3j–m and 4c). We first assessed the appetitive or aversive effects of optical pacing using a real-time place-preference (RTPP) assay (Fig. 2c). Mice spent

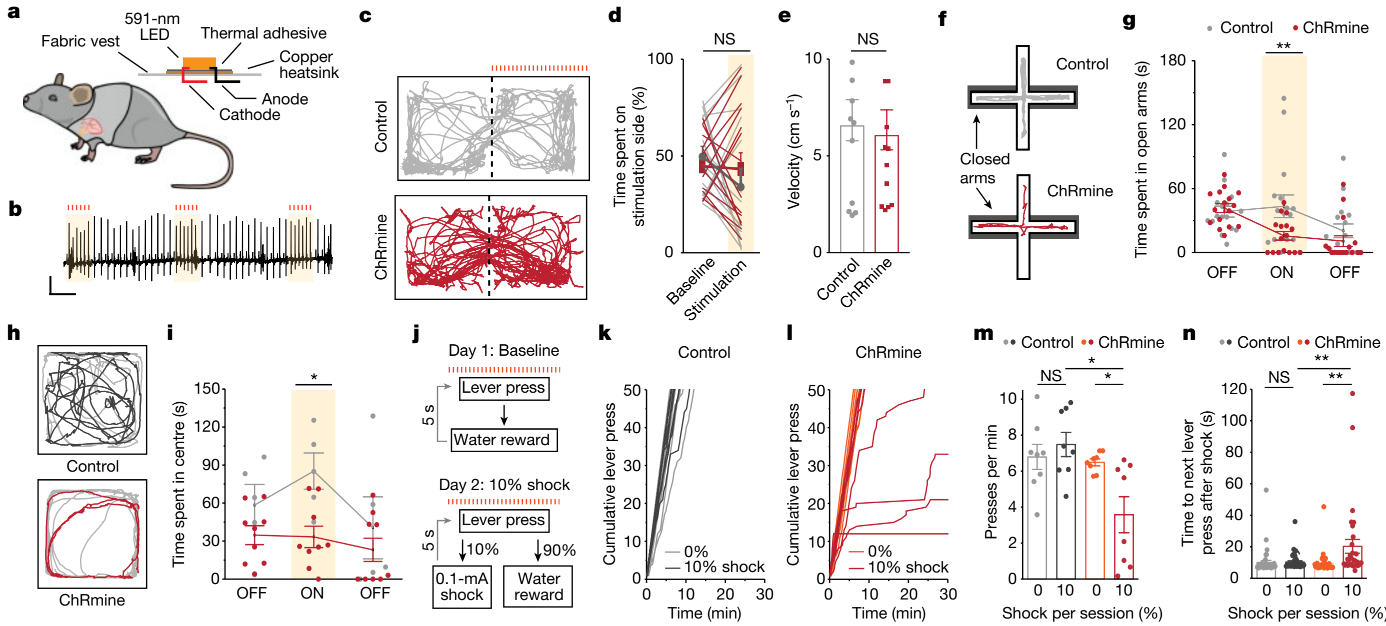


Fig. 2 | Optically induced tachycardia increases anxiety-like behaviour. **a**, Schematic of a micro-LED mounted to a wearable vest and fastened onto a mouse. **b**, Representative ECG trace of optically induced tachycardia (900 bpm for 500 ms every 2 s) used for all behavioural assays. Scale bar, 0.2 mV, 500 ms. **c**, Example path trace of a mouse with (red) or without (grey) ChRmine expression during an RTPP test, in which mice received optically induced cardiac pacing on one side of the chamber. **d**, Percentage of time spent on stimulation side during baseline and stimulation days for control (grey) and ChRmine-expressing (red) mice ($n = 16$ mice per group; two-way repeated-measures ANOVA with Bonferroni post hoc test: group (opsin) × time interaction $F_{(1,30)} = 2.29$, $P = 0.14$; group (opsin) effect $F_{(1,30)} = 6.2 \times 10^{-4}$, $P = 0.98$; time effect $F_{(1,30)} = 2.06$, $P = 0.16$. Bonferroni post hoc: control versus ChRmine $P = 0.71$ (baseline day); $P = 0.67$ (stimulation day)). NS, not significant. **e**, Average velocity on the optically paced side during RTPP ($n = 16$ mice per group; unpaired two-tailed *t*-test, $P = 0.81$). **f**, Example path trace of control (grey) and ChRmine-expressing (red) mice during an EPM test with optical pacing during the 5-min ON epoch of a 15-min trial. Open arms are vertical; closed arms are horizontal and bordered in grey. **g**, Time spent in open arms during 5-min epochs of EPM exploration ($n = 16$ mice per group; two-way repeated-measures ANOVA with Bonferroni post hoc test: group (opsin) × time interaction $F_{(2,60)} = 3.906$, $P = 0.0254$; group (opsin) effect $F_{(1,30)} = 3.297$, $P = 0.0794$; time effect $F_{(2,60)} = 9.75$, $P = 0.0002$. Bonferroni post hoc: ON epoch ChRmine versus control, $^{**}P = 0.0079$). **h**, Example path trace of control (grey) and ChRmine-expressing (red) mice during an OFT with optical pacing during the 3-min ON epoch of a 9-min trial. **i**, Time spent in the centre during 3-min epochs of OFT exploration ($n = 5$ (control), 9 (ChRmine) mice; two-way repeated-measures ANOVA with Bonferroni post hoc test: group (opsin) × time interaction $F_{(2,24)} = 1.531$, $P = 0.024$; group (opsin) effect $F_{(1,12)} = 5.69$, $P = 0.0035$; time effect $F_{(2,24)} = 3.42$, $P = 0.049$. Bonferroni post hoc: ON epoch ChRmine versus control, $^{*}P = 0.018$). **j**, Vogel conflict task to assay for cardiogenic effects on operant behaviour. Water-restricted mice were first trained for 2–3 weeks until each mouse was able to complete the 50 water-reward lever-press trials over 30 min for at least 3 consecutive days. On the day of the behavioural task, mice received optical pacing while completing a total of 50 lever-press trials per session (day 1). On the subsequent day, a 10% pseudorandom chance of shock was introduced upon lever press (day 2). **k**,**l**, Cumulative lever presses during 0% (day 1) and 10% shock (day 2) sessions for control (**k**) and ChRmine-expressing (**l**) mice ($n = 8$ mice). **m**, Average lever-pressing rate during 0% baseline or 10% shock trial sessions ($n = 8$ mice per group; two-way repeated-measures ANOVA with Bonferroni post hoc test: group (opsin) × condition (shock) interaction $F_{(1,14)} = 8.326$, $P = 0.0120$; group (opsin) effect $F_{(1,14)} = 7.39$, $P = 0.0166$; condition (shock) effect $F_{(1,14)} = 3.162$, $P = 0.0971$. Bonferroni post hoc: control 0% versus 10%, $P = 0.8933$; ChRmine 0% versus 10%, $^{*}P = 0.0106$; 0% control versus ChRmine, $P > 0.9999$; 10% control versus ChRmine, $^{**}P = 0.0010$). **n**, Elapsed time between a lever press resulting in shock and a subsequent lever press as a measure of the mouse's apprehension state. On 10% shock trials, 3 out of 8 mice did not complete the trial, so the time to next lever press for some trials cannot be measured ($n = 40$, 40, 40 and 32 presses per group in control 0%, control 10%, ChRmine 0% and ChRmine 10%; two-way ANOVA with Bonferroni post hoc test: group (opsin) × condition (shock) interaction $F_{(1,148)} = 6.478$, $P = 0.0119$; group (opsin) effect $F_{(1,148)} = 5.041$, $P = 0.0262$; condition (shock) effect $F_{(1,148)} = 7.253$, $P = 0.0079$. Bonferroni post hoc: control 0% versus 10%, $P > 0.9999$; ChRmine 0% versus 10%, $^{**}P = 0.0026$; 0% control versus ChRmine, $P > 0.9999$; 10% control versus ChRmine, $^{**}P = 0.0074$). Data are mean ± s.e.m.

an equal proportion of time on the paced and non-paced sides of the two-chamber arena, and showed no difference in locomotion compared to littermate controls—revealing that optically induced intermittent tachycardia was not intrinsically aversive and did not cause locomotor impairment (Fig. 2d,e). Optical pacing also did not modulate pain perception during a hot-plate test, with paced mice exhibiting comparable behavioural responses to control mice (Extended Data Fig. 5a–c).

Anxiety-like state evoked by cardiac pacing

By contrast, when we tested for anxiety-related behaviour using an elevated plus maze (EPM) assay, the same mice exhibited limited exploration of the open (exposed) arms of the apparatus after optical pacing, compared to control mice, preferring to remain within the protected areas of the closed arms (Fig. 2f,g). Paced mice also avoided the centre area during an open field test (OFT) (Fig. 2h,i). We observed no effects from illumination alone in control (opsin-negative) mice, and baseline anxiety levels between control and virally transduced groups were similar (Extended Data Fig. 5d–h). Increased anxiety-like behaviour induced by optical pacing during the EPM and OFT assays was similarly observed in female cohorts (Extended Data Fig. 5i–l). We found that mice that received intermittent cardiac pacing within baseline ranges (660 bpm) rather than elevated (900 bpm) ranges did not exhibit behavioural differences compared to control mice during the EPM or OFT (Extended Data Fig. 5m–p). Because continuous ventricular pacing can have a long-lasting effect on animal health[20,21], we also assessed for potential changes in baseline anxiety levels and

mobility in mice that were subjected to longer-term treatments of intermittent tachycardia (one-hour sessions every other day for two weeks) and did not observe locomotor or behavioural differences in these mice when compared to control mice during the EPM and OFT (Extended Data Fig. 5q–u).

We further investigated whether this context-dependent enhancement of anxiety-related behaviour could translate to a classical operant task, by using a trial-based variation of the Vogel conflict task in which water-restricted mice show willingness to seek a water reward even when the reward is coupled to a risk of mild shock[37] (Fig. 2j). Mice that received cardiac pacing performed similarly to control littermates when allowed to freely press for water with no delivery of the aversive stimulus (Fig. 2k–n). However, when random shocks were introduced in 10% of trials, optically paced mice were found to suppress or terminate water-seeking altogether (Fig. 2l,m). These mice exhibited increased apprehension, as revealed both by an overall decreased lever-pressing rate and by an increased time to the next subsequent lever press after a shock trial—consistent with the heightened levels of anxiety that were observed during the EPM and OFT (Fig. 2m,n). By contrast, control mice exhibited reduced water-seeking only when the frequency of shock trials was increased to 30% (Extended Data Fig. 6). This context-dependent influence of cardiac pacing on anxiety-like behaviour suggested that higher-order brain function was involved in the processing of interoceptive cues.

Optical pacing increases insula activity

We therefore next used the optical pacemaker to identify potential neural correlates and mechanisms of this observed behaviour along the heart–brain axis. First, transgenic TRAP2 mice, in which neurons with increased expression of the immediate early gene *Fos* can be labelled with tdTomato as a marker for neural activation[38], were used to perform a brain-wide screen to identify regions that were affected by optical pacing (Fig. 3a,b and Extended Data Fig. 7). Whole-brain tissue clearing[39], registration to a reference brain atlas[40] and automated cell counting together revealed that a number of brain regions exhibited increased expression of tdTomato in optically paced mice. These included areas associated with the central autonomic network[41,42], such as the insular cortex (including its visceral area (VISC), gustatory area (GU) and agranular insular area (AI)), prefrontal cortex (including the infralimbic area (ILA), prelimbic area (PL) and anterior cingulate area (ACA)) and brainstem (including the pons (P) and medulla (MY)) (Fig. 3b). Meanwhile, pointing to specificity, many other cortical regions that are not known to be involved in autonomic or interoceptive processing were not significantly activated, including primary sensory auditory (AUD) and visual (VIS) cortical areas, as well as the cerebellar vermis (VERM) and cerebellar nuclei (CBN) (Fig. 3b). Consistent with the TRAP2 mapping results, optical pacing similarly increased the endogenous expression of *Fos* mRNA in the posterior insular cortex (pIC) and in the brainstem (Fig. 3c and Extended Data Fig. 8). In particular, sensory relay circuits of the nucleus tractus solitarius (NTS), as well as noradrenergic neurons in the locus coeruleus (LC), which are involved in arousal and stress[43], also exhibited prominent *Fos* labelling (Extended Data Fig. 8).

We next investigated the cardiac-pacing-induced neural dynamics at single-neuron resolution using in vivo electrophysiology in awake mice (Fig. 3d–h). On the basis of the observed increased *Fos* expression in the pIC after cardiac pacing, and because the insular cortex is a key cortical hub for interoception[44], we used four-shank Neuropixels 2.0 probes to obtain rich multi-regional recordings of the pIC and surrounding regions (Fig. 3d,e). Notably, we observed pacing-evoked increases in insular activity at both the single-unit and population levels (Fig. 3f,g), whereby pIC neurons were acutely activated by cardiac stimulation with heterogeneous temporal dynamics and returned to basal levels of activity after pacing offset. By contrast, no significant changes in activity were recorded in control mice during photostimulation, ruling out potential light- or heat-induced representations in this region (Fig. 3g). We also observed a substantial diversity in temporal dynamics triggered by pacing in other recorded regions, including acute responses in the somatosensory cortex and delayed responses in the striatum during stimulation offset (Fig. 3h). Our data show that the pIC and other regions of the central autonomic network are distinctly engaged by optically evoked tachycardia[41,42], and are in line with human neuroimaging studies that have correlated these brain areas with cardiac interoception[45–47].

Insular inhibition reduces cardiac anxiety effects

To determine whether the anxiogenic circuitry recruited by cardiac pacing could be specifically modulated to influence behaviour, we next performed optogenetic inhibition with the 473 nm (blue light)-activated inhibitory channelrhodopsin iC++. We targeted the pIC (with well-established roles in both processing and regulating cardiac sensory signals and anxiety-related behaviours[45–48]) and the medial prefrontal cortex (mPFC) (crucially involved in reward and aversion processing, as well as associated with cardiovascular arousal[49]). To perform simultaneous optogenetic inhibition of the cortex and optically paced tachycardia, we bilaterally injected AAVdj-hSyn::iC++-eYFP or AAVdj-hSyn::eYFP control virus and implanted fibre-optic cannulas into the pIC or mPFC of mice expressing ChRmine in the heart[50] (Fig. 4a). As expected, in eYFP-expressing control mice, we found that 473-nm illumination of the pIC did not affect the reduction of water-seeking by optical pacing in trials with a risk of shock (Fig. 4b,c). By contrast, the same intervention in mice expressing iC++ instead of eYFP alone in pIC did reverse the reduction of water-seeking (Fig. 4d–g); all mice receiving pIC inhibition completed the water-retrieving task and exhibited decreased apprehension, with the time to next lever press after shock reduced to near baseline levels (Fig. 4g). Similarly, iC++ inhibition increased open-arm exploration time during the EPM assay in optically paced mice relative to eYFP controls (Fig. 4h).

The attenuation of the anxiogenic effect of optical pacing exhibited specificity to pIC inhibition; inhibition of the mPFC did not decrease cardiac-associated anxiogenic behaviours relative to eYFP controls (Fig. 4i–n). To test for any direct anxiolysis from optogenetic inhibition of the pIC (that is, not through modulation of the cardiac-pacing effect), we performed the EPM assay and the lever-pressing task in the absence of cardiac pacing and with 30% shock trials to allow the detection of pacing-independent apprehensive behaviour (Extended Data Fig. 9a–d). Inhibition of the pIC without cardiac pacing did not increase open-arm exploration, affect lever-press suppression or influence heart rate (Extended Data Fig. 9). Thus, pIC inhibition alone appeared to be insufficient to induce anxiolysis, consistent with previous reports[48]. Together, these results are in line with a model in which the pIC is important for mediating the anxiety-related and apprehensive behaviours that arise from direct cardiac pacing.

Discussion

In this study, we have developed a method for noninvasive optogenetic control of specific cardiac rhythms during active behaviour. We show that the optically induced tachycardia was not intrinsically aversive, but rather elicited anxiety-like behaviours and apprehension in potentially risky environments. Although diverse mechanisms may contribute to this effect, we consider that anxiogenic effects of evoked tachycardia are not likely to be mediated through a reduction in blood pressure[51], as drugs that reduce systolic blood pressure tend to be anxiolytic (for example, propranolol and clonidine) or neutral (for example, Ca^{2+}-channel blockers). Our observation of anxiogenesis in response to increased heart rates (900 bpm or 15 Hz) is in line with clinical observations that accelerated heart rates—but not other forms of

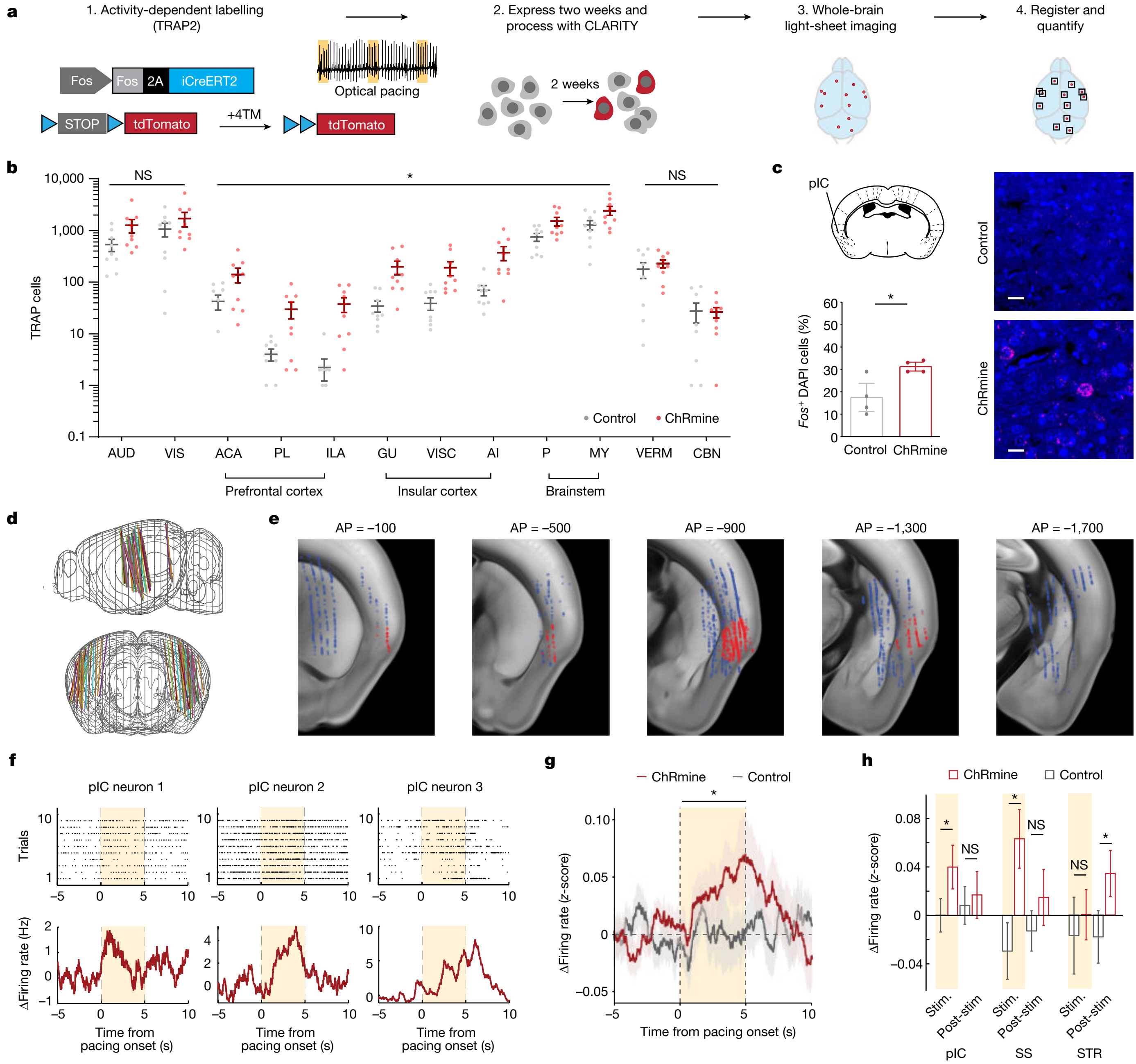


Fig. 3 | Whole-brain screen for regions that are activated by optically induced tachycardia. a, Schematic for whole-brain activity mapping to identify regions that are activated by optically paced tachycardia. Double-transgenic TRAP2;Ai14 reporter mice were injected with 4-hydroxytamoxifen (4TM) and treated with optically induced tachycardia for 15 min. After two weeks of tdTomato reporter gene expression, mice were euthanized and processed with CLARITY. Whole brains were imaged with a light-sheet microscope, followed by automated registration to a common brain atlas, cell segmentation and quantification to identify brain regions with differential accumulation of activated TRAP (tdTomato$^+$) cells. **b**, Regional cell counts of paced (ChRmine, red) versus control (grey) cohorts sorted from anterior to posterior anatomical regions with select regions from the central autonomic network with increased TRAP cells and regions outside of the central autonomic network without statistical significance ($n = 9$ per group; multiple two-sided t-tests corrected for multiple comparisons with the Benjamini and Hochberg method (*false discovery rate (FDR) = 10%)). Significantly activated regions include the prefrontal cortex (ACA, PL and ILA), insular cortex (GUI, VISC and AI) and brainstem (P and MY). Non-significant regions include primary sensory cortices (AUD, VIS) and the cerebellum (VERM, CBN). **c**, Percentage of cells that are *Fos*$^+$, determined from in situ hybridization for *Fos* mRNA (magenta) after cardiac pacing in control and ChRmine-expressing mice in the pIC ($n = 4$ mice per group; unpaired two-tailed t-test, *$P = 0.020$). Scale bars, 20 μm. **d**, Electrode tracks from $n = 5$ mice (3 ChRmine and 2 control) over 60 recording sessions co-registered to the common Allen Brain Atlas. **e**, Locations of recorded single units overlaid onto the Allen Brain Atlas. Red denotes units in the insular cortex. AP, anterior–posterior. **f**, Spike raster and changes in firing rate for three example insular neurons after 900-bpm pacing. **g**, Population-averaged changes in firing rate of insular neurons from ChRmine (red, $n = 391$) or control (grey, $n = 228$) mice (one-sided P values from hierarchical bootstrap: $P = 0.026$ (during 5 s pacing); $P = 0.357$ (during 5 s after pacing)). **h**, Average change in baseline firing rate per brain region across 5-s epochs during and after photostimulation in control (grey) and ChRmine-expressing (red) mice. Single units were obtained from the pIC ($n = 228$ (control), $n = 391$ (ChRmine)); somatosensory cortex (SS; $n = 77$ (control), $n = 368$ (ChRmine)); and striatum (STR; $n = 70$ (control), $n = 800$ (ChRmine)). One-sided P values from hierarchical bootstrap: pIC: $P = 0.026$ (stimulation; stim.), $P = 0.36$ (post-stimulation; post-stim.); SS: $P = 0.0014$ (stim.), $P = 0.16$ (post-stim.); STR: $P = 0.29$ (stim.), $P = 0.036$ (post-stim.). Data are mean ± s.e.m.

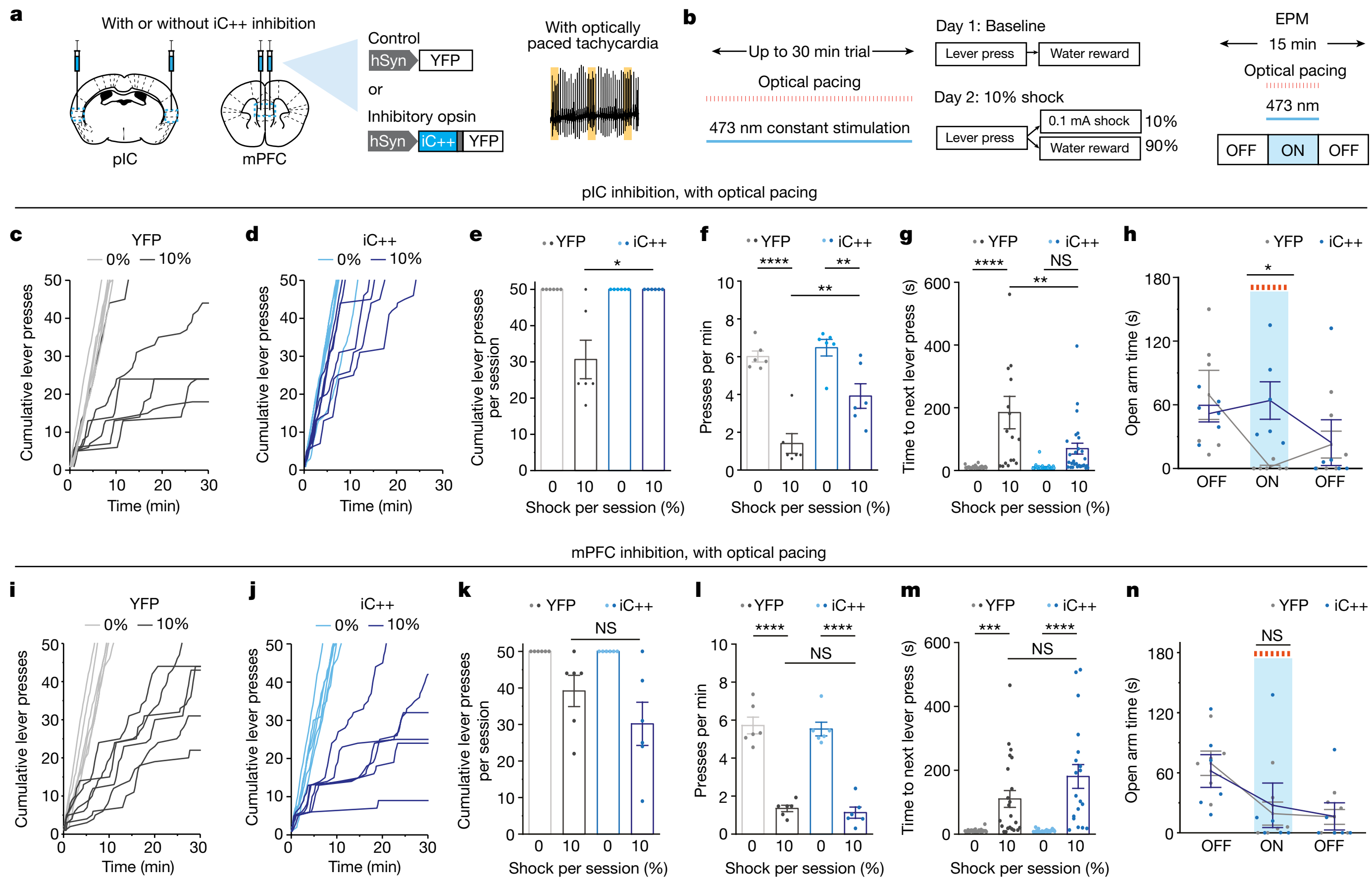


Fig. 4 | Optogenetic inhibition of the posterior insula attenuates the anxiogenic response from optical pacing. **a**, Illustration of the experimental protocol for simultaneous optically induced tachycardia and optogenetic inhibition of the pIC or mPFC using AAVdj-hSyn::iC++-YFP or control virus (YFP only). **b**, Left, conditions for the Vogel conflict task, in which mice received both 473-nm constant illumination in the pIC or mPFC and optically induced tachycardia during the behavioural task. Right, illustration of the experimental protocol for simultaneous optogenetic inhibition of the pIC with optical pacing during the EPM test. **c**,**d**, Cumulative lever presses during baseline (day 1) and 10% shock (day 2) sessions for mice expressing control (YFP) (**c**) or iC++ (**d**) in the pIC with optical pacing ($n = 6$ mice per group). **e**, Cumulative number of lever presses completed in each session. Owing to increased apprehension, only 1 out of 6 control mice completed the 50-lever-press session on the 10% shock session ($n = 6$ per group; two-sided Wilcoxon rank-sum test, $^{*}P = 0.0152$). **f**, Average lever-pressing rate for 0% and 10% shock experimental sessions. Note that iC++ inhibition partially restores overall rates of lever pressing, but not to baseline levels ($n = 6$ per group; two-way ANOVA with Bonferroni post hoc test: group × condition interaction $F_{(1,10)} = 5.533$, $P = 0.0405$; group (opsin) effect $F_{(1,10)} = 7.439$, $P = 0.0213$; condition (shock) effect $F_{(1,10)} = 67.8$, $P < 0.0001$. Bonferroni post hoc: 0% shock YFP versus iC++, $P > 0.9999$; 10% shock YFP versus iC++, $^{**}P = 0.0036$; YFP 0% versus 10% shock, $^{****}P < 0.0001$; iC++ 0% versus 10% shock, $^{**}P = 0.0039$). **g**, Time to next lever press after shock. Note that iC++ inhibition reduces apprehension to no-shock levels ($n = 30$, 17, 30 and 30 presses per group in YFP 0%, YFP 10%, iC++ 0% and iC++ 10%; two-way ANOVA with Bonferroni post hoc test: group (opsin) × condition (shock) interaction $F_{(1,103)} = 8.7$, $P = 0.0039$; group (opsin) effect $F_{(1,103)} = 8.6$, $P = 0.0041$; condition (shock) effect $F_{(1,103)} = 35.6$, $P < 0.0001$. Bonferroni post hoc: YFP 0% versus 10%, $^{****}P < 0.0001$; iC++ 0% versus 10%, $P = 0.099$; 0% YFP versus iC++, $P > 0.9999$; 10% YFP versus iC++, $^{**}P = 0.0011$). **h**, Time spent in open arms during 5-min epochs of EPM exploration ($n = 6$ per group; two-way repeated-measures ANOVA with Bonferroni post hoc test: group (opsin) × time interaction $F_{(2,20)} = 3.543$, $P = 0.0482$; group (opsin) effect $F_{(1,10)} = 1.251$, $P = 0.2894$; time effect $F_{(2,20)} = 3.058$, $P = 0.0694$. Bonferroni post hoc: ON epoch YFP versus iC++, $^{*}P = 0.0323$). **i**,**j**, Cumulative lever presses with the same conditions as **c**,**d** but with expression of YFP (**i**) or iC++ (**j**) in the mPFC ($n = 6$ mice). **k**, Cumulative number of lever presses during each session. Note that in 10% shock sessions, both YFP- and iC++-expressing mPFC mice cease lever pressing ($n = 6$ per group; two-sided Wilcoxon rank-sum test, $P = 0.2987$). **l**, Average lever-pressing rate for 0% and 10% shock sessions ($n = 6$ per group; two-way repeated-measures ANOVA with Bonferroni post hoc test: group × condition interaction $F_{(1,10)} = 0.002521$, $P = 0.9609$; group (opsin) effect $F_{(1,10)} = 0.4370$, $P = 0.5235$; condition (shock) effect $F_{(1,10)} = 154.1$, $P < 0.0001$. Bonferroni post hoc: 0% shock YFP versus iC++, $P > 0.9999$; 10% shock YFP versus iC++, $P > 0.9999$; YFP 0% versus 10% shock, $^{****}P = 1.1 \times 10^{-5}$; iC++ 0% versus 10% shock, $^{****}P = 1.0 \times 10^{-6}$). **m**, Time to next lever press after shock ($n = 30$, 22, 30 and 19 presses per group in mPFC YFP 0%, YFP 10%, iC++ 0% and iC++ 10%; two-way ANOVA with Bonferroni post hoc test: group (opsin) × condition (shock) interaction $F_{(1,97)} = 3.703$, $P = 0.05725$; group (opsin) effect $F_{(1,97)} = 3.610$, $P = 0.0604$; condition (shock) effect $F_{(1,97)} = 54.18$, $P < 0.000001$. Bonferroni post hoc: YFP 0% versus 10%, $^{***}P = 0.0009$; iC++ 0% versus 10%, $^{****}P = 2.95 \times 10^{-8}$; 0% YFP versus iC++, $P > 0.9999$; 10% YFP versus iC++, $P = 0.0698$). **n**, Time spent in open arms during 5-min epochs of EPM exploration with (iC++, blue) and without (YFP, grey) mPFC inhibition ($n = 6$ per group; two-way repeated-measures ANOVA with Bonferroni post hoc test: group (opsin) × time interaction $F_{(2,20)} = 0.3929$, $P = 0.6802$; group (opsin) effect $F_{(1,10)} = 0.00039$, $P = 0.9846$; time effect $F_{(2,20)} = 17.41$, $P < 0.0001$. Bonferroni post hoc: ON epoch YFP versus iC++, $P > 0.9999$).

altered haemodynamics (for example, increased heart rate variability)—are associated with panic and other anxiety-related disorders[52,53]. The altered rate, rather than the external nature of cardiac contraction timing, appears to be important; for example, we found that intermittent or asynchronous stimulation close to baseline heart rates at 660 bpm (11 Hz) did not result in anxiety-like behaviour.

In further investigations of the mechanisms that underlie these behaviours, we found that optogenetic pacing activated the pIC,

consistent with studies of cardiovascular control in anaesthetized rodents[54] and neuroimaging studies of cardiac interoception and reflex control in humans[45–47], including in the setting of panic and anxiety[49,55]. It remains unclear whether this pathway can be modulated by peripheral baroreceptive sensory neurons or other sensory mechanisms that detect changes in blood pressure[56–58], and it is possible that there are additional ways in which cardiac viscerosensory information can be relayed to higher cortical areas[41,42]. The anxiogenic behavioural effects of cardiac pacing were attenuated during optogenetic inhibition of the pIC, suggesting that the insula has a causal role in integrating sensory information from the heart with a contextual assessment of environmental risk to produce adaptive behavioural patterns. Our findings support the idea that the insular cortex is involved in monitoring not only consummatory[59–61] but also entirely internal interoceptive states to instruct relevant behavioural responses, as predicted from human neuroimaging studies of cardiac interoception[45–47].

This study shows that cell-type-specific, temporally precise, noninvasive perturbation of organ-scale physiology is possible in fully intact, freely behaving mammals. Although we have applied our approach mainly to study animal behaviour over a period of minutes, future integration with miniaturized wireless devices[15] may facilitate longer-term studies to modulate targeted populations of cells over days to weeks while alleviating the need for intimate contact with the light source. Furthermore, refinement in cell-type-targeting strategies may enable minimally invasive to noninvasive optogenetic dissection of specific cell types (for example, pacemaker cells[28], Purkinje fibres[27] and cardiac ganglions[12]) to determine their effects on regulating cardiac electrophysiology and behaviour. Finally, our approach, which requires no specialized optoelectronics or surgery, has the potential for broad application to a range of physiological systems throughout the body—opening up numerous opportunities to explore the complex interactions between physiological systems in health, disease and treatment.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-05748-8.

1. Critchley, H. D. & Harrison, N. A. Visceral influences on brain and behavior. *Neuron* **77**, 624–638 (2013).
2. Chen, W. G. et al. The emerging science of interoception: sensing, integrating, interpreting, and regulating signals within the self. *Trends Neurosci.* **44**, 3–16 (2021).
3. Craig, A. D. How do you feel? Interoception: the sense of the physiological condition of the body. *Nat. Rev. Neurosci.* **3**, 655–666 (2002).
4. Critchley, H. D. & Garfinkel, S. N. Interoception and emotion. *Curr. Opin. Psychol.* **17**, 7–14 (2017).
5. Tural, U. & Iosifescu, D. V. The prevalence of mitral valve prolapse in panic disorder: a meta-analysis. *Psychosomatics* **60**, 393–401 (2019).
6. Frommeyer, G., Eckardt, L. & Breithardt, G. Panic attacks and supraventricular tachycardias: the chicken or the egg? *Neth. Heart J.* **21**, 74–77 (2013).
7. Oppenheimer, S. & Cechetto, D. The insular cortex and the regulation of cardiac function. *Compr. Physiol.* **6**, 1081–1133 (2011).
8. Schandry, R. Heart beat perception and emotional experience. *Psychophysiology* **18**, 483–488 (1981).
9. Lange, C. G. & James, W. *The Emotions* (Williams & Wilkins, 1922).
10. Gogolla, N. The insular cortex. *Curr. Biol.* **27**, R580–R586 (2017).
11. Klein, A. S., Dolensek, N., Weiand, C. & Gogolla, N. Fear balance is maintained by bodily feedback to the insular cortex in mice. *Science* **374**, 1010–1015 (2021).
12. Rajendran, P. S. et al. Identification of peripheral neural circuits that regulate heart rate using optogenetic and viral vector strategies. *Nat. Commun.* **10**, 1944 (2019).
13. Ressler, K. J. & Mayberg, H. S. Targeting abnormal neural circuits in mood and anxiety disorders: from the laboratory to the clinic. *Nat. Neurosci.* **10**, 1116–1124 (2007).
14. Entcheva, E. & Kay, M. W. Cardiac optogenetics: a decade of enlightenment. *Nat. Rev. Cardiol.* **18**, 349–367 (2021).
15. Won, S. M., Song, E., Reeder, J. T. & Rogers, J. A. Emerging modalities and implantable technologies for neuromodulation. *Cell* **181**, 115–135 (2020).
16. Shahriari, D., Rosenfeld, D. & Anikeeva, P. Emerging frontier of peripheral nerve and organ interfaces. *Neuron* **108**, 270–285 (2020).
17. Montgomery, K. L., Iyer, S. M., Christensen, A. J., Deisseroth, K. & Delp, S. L. Beyond the brain: optogenetic control in the spinal cord and peripheral nervous system. *Sci. Transl. Med.* **8**, 337rv5 (2016).
18. Berntson, G. G. & Khalsa, S. S. Neural circuits of interoception. *Trends Neurosci.* **44**, 17–28 (2021).
19. Mickle, A. D. et al. A wireless closed-loop system for optogenetic peripheral neuromodulation. *Nature* **565**, 361–365 (2019).
20. Hulsmans, M. et al. A miniaturized, programmable pacemaker for long-term studies in the mouse. *Circ. Res.* **123**, 1208–1219 (2018).
21. Bilchick, K. C. et al. Differential regional gene expression from cardiac dyssynchrony induced by chronic right ventricular free wall pacing in the mouse. *Physiolog. Genomics* **26**, 109–115 (2006).
22. Laughner, J. I. et al. A fully implantable pacemaker for the mouse: from battery to wireless power. *PLoS ONE* **8**, e76291 (2013).
23. Gutruf, P. et al. Wireless, battery-free, fully implantable multimodal and multisite pacemakers for applications in small animal models. *Nat. Commun.* **10**, 5742 (2019).
24. Nussinovitch, U. & Gepstein, L. Optogenetics for in vivo cardiac pacing and resynchronization therapies. *Nat. Biotechnol.* **33**, 750–754 (2015).
25. Nyns, E. C. A. et al. Optogenetic termination of ventricular arrhythmias in the whole heart: towards biological cardiac rhythm management. *Eur. Heart J.* **38**, 2132–2136 (2016).
26. Vogt, C. C. et al. Systemic gene transfer enables optogenetic pacing of mouse hearts. *Cardiovasc. Res.* **106**, 338–343 (2015).
27. Zaglia, T. et al. Optogenetic determination of the myocardial requirements for extrasystoles by cell type-specific targeting of ChannelRhodopsin-2. *Proc. Natl Acad. Sci. USA* **112**, E4495–E4504 (2015).
28. Lee, F. K. et al. Genetically engineered mice for combinatorial cardiovascular optobiology. *eLife* **10**, e67858 (2021).
29. Marshel, J. H. et al. Cortical layer-specific critical dynamics triggering perception. *Science* **365**, eaaw5202 (2019).
30. Chen, R. et al. Deep brain optogenetics without intracranial surgery. *Nat. Biotechnol.* **39**, 161–164 (2021).
31. Pacak, C. A. et al. Recombinant adeno-associated virus serotype 9 leads to preferential cardiac transduction in vivo. *Circ. Res.* **99**, e3–e9 (2006).
32. Bedbrook, C. N. et al. Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. *Nat. Methods* **16**, 1176–1184 (2019).
33. Kishi, K. E. et al. Structural basis for channel conduction in the pump-like channelrhodopsin ChRmine. *Cell* **185**, 672–689 (2022).
34. Sgoifo, A., Carnevali, L., Pico Alfonso, M. D. L. A. & Amore, M. Autonomic dysfunction and heart rate variability in depression. *Stress* **18**, 343–352 (2015).
35. Lampert, R. et al. Emotional and physical precipitants of ventricular arrhythmia. *Circulation* **106**, 1800–1805 (2002).
36. Taggart, P., Critchley, H. & Lambiase, P. Heart–brain interactions in cardiac arrhythmia. *Heart* **97**, 698–708 (2011).
37. Kim, C. K. et al. Molecular and circuit-dynamical identification of top-down neural mechanisms for restraint of reward seeking. *Cell* **170**, 1013–1027 (2017).
38. Allen, W. E. et al. Thirst-associated preoptic neurons encode an aversive motivational drive. *Science* **357**, 1149 (2017).
39. Ye, L. et al. Wiring and molecular features of prefrontal ensembles representing distinct experiences. *Cell* **165**, 1776–1788 (2016).
40. Goubran, M. et al. Multimodal image registration and connectivity analysis for integration of connectomic data from microscopy to MRI. *Nat. Commun.* **10**, 5504 (2019).
41. Benarroch, E. E. The central autonomic network: functional organization, dysfunction, and perspective. *Mayo Clin. Proc.* **68**, 988–1001 (1993).
42. Azzalini, D., Rebollo, I. & Tallon-Baudry, C. Visceral signals shape brain dynamics and cognition. *Trends Cogn. Sci.* **23**, 488–509 (2019).
43. Carter, M. E. et al. Tuning arousal with optogenetic modulation of locus coeruleus neurons. *Nat. Neurosci.* **13**, 1526 (2010).
44. Livneh, Y. & Andermann, M. L. Cellular activity in insular cortex across seconds to hours: sensations and predictions of bodily states. *Neuron* **109**, 3576–3593 (2021).
45. Critchley, H. D., Wiens, S., Rotshtein, P., Öhman, A. & Dolan, R. J. Neural systems supporting interoceptive awareness. *Nat. Neurosci.* **7**, 189–195 (2004).
46. Khalsa, S. S., Rudrauf, D., Feinstein, J. S. & Tranel, D. The pathways of interoceptive awareness. *Nat. Neurosci.* **12**, 1494–1496 (2009).
47. Hassanpour, M. S. et al. The insular cortex dynamically maps changes in cardiorespiratory interoception. *Neuropsychopharmacology* **43**, 426–434 (2018).
48. Gehrlach, D. A. et al. Aversive state processing in the posterior insular cortex. *Nat. Neurosci.* **22**, 1424–1437 (2019).
49. Pollatos, O., Schandry, R., Auer, D. P. & Kaufmann, C. Brain structures mediating cardiovascular arousal and interoceptive awareness. *Brain Res.* **1141**, 178–187 (2007).
50. Berndt, A. et al. Structural foundations of optogenetics: determinants of channelrhodopsin ion selectivity. *Proc. Natl Acad. Sci. USA* **113**, 822–829 (2016).
51. Carnovale, C. et al. Antihypertensive drugs and brain function: mechanisms underlying therapeutically beneficial and harmful neuropsychiatric effects. *Cardiovasc. Res.* **2022**, cvac110 (2022).
52. Gorman, J. M. & Sloan, R. P. Heart rate variability in depressive and anxiety disorders. *Am. Heart J.* **140**, S77–S83 (2000).
53. Chalmers, J. A., Quintana, D. S., Abbott, M. J.-A. & Kemp, A. H. Anxiety disorders are associated with reduced heart rate variability: a meta-analysis. *Front. Psychiatry* **5**, 80 (2014).
54. Oppenheimer, S. M. & Cechetto, D. F. Cardiac chronotropic organization of the rat insular cortex. *Brain Res.* **533**, 66–72 (1990).
55. Schulz, S. M. Neural correlates of heart-focused interoception: a functional magnetic resonance imaging meta-analysis. *Phil. Trans. R. Soc. B* **371**, 20160018 (2016).
56. Zeng, W.-Z. et al. PIEZOs mediate neuronal sensing of blood pressure and the baroreceptor reflex. *Science* **362**, 464 (2018).
57. Min, S. et al. Arterial baroreceptors sense blood pressure through decorated aortic claws. *Cell Rep.* **29**, 2192–2201 (2019).

58. Zhang, Z. & Oppenheimer, S. M. Characterization, distribution and lateralization of baroreceptor-related neurons in the rat insular cortex. *Brain Res.* **760**, 243–250 (1997).
59. Livneh, Y. et al. Homeostatic circuits selectively gate food cue responses in insular cortex. *Nature* **546**, 611–616 (2017).
60. Kusumoto-Yoshida, I., Liu, H., Chen, B. T., Fontanini, A. & Bonci, A. Central role for the insular cortex in mediating conditioned responses to anticipatory cues. *Proc. Natl Acad. Sci. USA* **112**, 1190 (2015).
61. Wu, Y. et al. The anterior insular cortex unilaterally controls feeding in response to aversive visceral stimuli in mice. *Nat. Commun.* **11**, 640 (2020).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

Methods

Mice

All animal procedures followed animal care guidelines approved by Stanford University's Administrative Panel on Laboratory Animal Care (APLAC) and guidelines of the National Institutes of Health. Investigators were not blinded to the genotypes of the mice. Male and female wild-type C57BL6/J (JAX 0064) mice were used for most behavioural experiments unless specified otherwise, and all mice were 8–12 weeks old at the time of starting behavioural experiments. Mice were housed in plastic cages with disposable bedding on a standard light cycle with food and water available ad libitum, except when placed on water restriction. When on water restriction, mice were provided with 1 ml of water each day and maintained above 85% of baseline weight. Behavioural experiments were performed during the dark phase.

Molecular cloning

A 685-bp fragment containing the promoter region of the mouse troponin gene was amplified from a wild-type mouse using CGCACGCGTGAGGCCATTTGGCTCATGAGAAGC and CATGGATCCTCTAGAAAGGGCCATGGATTTCCTG primers, cloned upstream of ChRmine-p2A-oScarlet using MluI and BamHI sites in an AAV backbone, sequence-verified and tested for expression in dissociated neonatal cardiomyocytes.

In vitro cardiomyocyte experiments

Dissociated neonatal mouse cardiomyocytes prepared using the Pierce Isolation Kit (Thermo Fisher Scientific, 88289) were transfected with rAAV-mTNT::ChRmine-p2A-oScarlet (1 µl of 8 × 12 viral genomes (vg) per ml in 500 µl of medium). Three to five days after infection, individual cardiomyocytes were identified under a light microscope. Optical stimulation was provided by a Spectra X Light engine at 585 nm (LumenCore) coupled to a Leica DM LFSA microscope and synchronized with video recording at 100 fps using LabView software. Laser power leaving the imaging objective was measured with an optical power meter (Thorlabs PM100D). Videos were analysed for contraction using custom scripts in MATLAB.

In vivo systemic viral delivery

Wild-type mice aged three to four weeks were anaesthetized with isoflurane and rAAV-mTNT::ChRmine-p2A-oScarlet (2×10^{11} vg per mouse) or vehicle was delivered by retro-orbital injection. Our selected titres were previously used for systemic viral transduction of ChR2 in the heart[26]. A total volume of 60 µl 0.9% NaCl saline solution was injected into the right retro-orbital sinus using a 28G needle, then allowed to recover on a warming pad before being returned to the home cage.

Optical pacemaker in vivo characterization

Mice were tested three weeks after injection of the pacemaker virus. Electrocardiography signals were collected using commercial instruments (Rodent Surgical Monitor+, Indus Instruments), with anaesthetized mice placed in a supine position and limbs placed in contact with electrode pads via a conductive gel. A 594-nm laser (LaserGlow) was attached to a fibre-optic patch cord (Thorlabs) terminating in a 200-µm-diameter, 0.39-NA fibre (Thorlabs) which was positioned against the chest. Optical power was adjusted using the laser's built-in power modulator and measured with an optical power meter (Thorlabs) at the fibre tip. Stimulation was performed with a pulse width of 10 ms and an inter-pulse interval ranging from 120 ms (equivalent to 500 bpm) to 67 ms (900 bpm), controlled by a TTL signal generator (Master-8). Heart rate (bpm) was derived from the heart rate interval between successive R waves (RR interval) obtained from ECG recordings. Fidelity of photoactivated QRS complexes was quantified by counting the number of beats at a set frequency divided by the number of total beats measured during the middle 20 s of a 30-s stimulation period.

Measurements of systolic blood pressure

Mice were anaesthetized (1.5–2% isofluorane) and placed in the supine position with the chest shaved. Systolic blood pressure measurements were performed using a 1.4-F pressure sensor mounted Millar catheter (SPR-671, ADInstruments) and recorded using LabChart 7 Pro (ADInstruments). The catheter was inserted via the right carotid artery into the left ventricle. A 589-nm laser was used to deliver 240 mW mm^{-2} light across intact chest at either constant or intermittent (500 ms ON, 1,500 ms OFF) optical stimulation at 900 bpm with a 10-ms pulse width for 30 s to assess optogenetic pacing effects on systolic blood pressure in real time.

Wearable optical pacemaker hardware

Custom-made wearable optical stimulators were constructed using 3 × 4.5 mm 591-nm PC Amber Rebel LEDs (Luxeon LXM2-PL01-0000). 30AWG flexible silicone wire (Striveday) was soldered to the LED pad and coated with electrically insulating, thermally conductive epoxy (Arctic Alumina), and adhered to copper sheet cut to 10 × 15 mm for thermal dissipation and subsequently glued to a fabric vest designed for freely moving mouse behaviour (Coulbourn A71-21M25). Wiring was held in place on the vest using hot glue and the free ends were inserted into a breadboard for stimulus control by an LED Driver (Thorlabs LEDD1B T-Cube). The optical power was set to 160–240 mW mm^{-2} measured from the surface of the LED. Light was delivered at intervals consisting of a 10-ms pulse width at 15 Hz (900 bpm) for 500 ms with 1,500 ms OFF time by using either a Master-8 or an Arduino microcontroller synchronized to behaviour recording software. Computer-aided design schematics were created with Onshape. Thermal measurements were performed using a FLIR C2 Compact thermal camera (FLIR) and the thermal profile at the surface of the micro-LED is plotted in Extended Data Fig. 4c.

Freely moving behaviour with pacemaker

All mice were habituated to the experimenter and handled for at least three days, and in addition allowed to acclimatize to wearing optical pacemaker hardware for at least five days, before behavioural experiments. Fur over the chest was removed (Nair) at least five days before behavioural experiments. Mice were briefly anaesthetized with isoflurane before the placement of the optical pacing vest and allowed to fully recover in the home cage (at least 1 h) before experiments. We used a stimulation protocol consisting of a 10-ms pulse width at 15 Hz (900 bpm) with 500 ms ON time and 1,500 ms OFF time to introduce intermittent tachycardia or 10-ms pulse width at 11 Hz (660 bpm) with a Poisson distribution to introduce increased heart rate variability. Mice received optical stimulation during the ON periods of the behavioural assay from the wearable micro-LED device in both control and ChRmine cohorts. No statistical difference in behaviour was observed between virally transduced and control groups at baseline, suggesting that there were no side effects from transgene delivery. No statistical difference in behaviour was observed in control groups before, during and after optical stimulation, suggesting that there were no effects from light delivery alone.

RTPP

Mice were placed in a custom-built RTPP chamber (30.5 × 70 cm) on day 1 to determine their baseline preference for each side of the chamber. Behavioural tracking was performed using blinded automated software (Noldus Ethovision). On day 2, mice were stimulated whenever they were on one side of the chamber. Stimulation sides were randomly assigned and counterbalanced across mice. Each session lasted 20 min.

EPM

The EPM was made of grey plastic (Med Associates). Mice were gently placed in the closed arm of the EPM. Mice were allowed to freely explore

the maze for a 5-min baseline 'off' period, followed by a 5-min 'on' period during which optical stimulation was delivered, and finally a 5-min 'off' period. Behavioural tracking was performed using blinded automated software (Noldus Ethovision) and the overall time spent in open arms was reported for each epoch.

OFT

Mice were placed in a 60 × 60-cm arena and allowed to freely explore during a 9-min session. Optical stimulation was delivered during the middle 3-min epoch. Movement was tracked with a video camera positioned above the arena. To assess anxiety-related behaviour, the chamber was divided into a peripheral and centre (48 × 48 cm) region.

Operant lever-pressing task

Water-restricted mice were trained to lever press for a small water reward (around 10 µl water) while freely moving in an operant condition box containing a single retractable lever and a shock grid floor (Coulbourn). Mice were allowed to retrieve a maximum of 50 rewards per day, and sessions were terminated after all rewards had been retrieved or after 30 min. After each lever press, the lever was retracted for 5 s before extending again. After mice retrieved 50 rewards for at least 3 consecutive days (typically 2–3 weeks of training), they were allowed to proceed with stimulation experiments. On shock days, mice were given a 1-s, 0.1-mA foot shock after 10% of lever presses instead of water. Shocks were delivered in a pseudorandom order on lever-press trials 5, 13, 24, 31 and 44, and the time to the next lever press was measured from the time elapsed for these trials until the subsequent lever press. During stimulation experiments (both baseline and shock days), optical stimulation was delivered throughout the experiment. Water was delivered using a custom set-up consisting of a lick spout (Popper and Sons, stainless steel 18-gauge) and a solenoid (Valcor, SV74P61T1) controlled by a microcontroller (Arduino Uno R3). Licking was monitored using a capacitive sensing board (Arduino Tinker Kit) wired to the lick spout and interfacing with the microcontroller. Shocks were delivered using an 8-pole scrambled shock floor (Coulbourn). Behavioural stimuli—lever presentations and retractions, and shocks—were controlled with Coulbourn Graphic State software. The timing of lever presses and licks was also recorded at 5 kHz using a data-acquisition hardware (National Instruments, NI PCIe-6343-X).

TRAP2 labelling

Fos$^{2A\text{-}iCreER}$ (TRAP2; JAX 030323) mice were backcrossed onto a C57BL6/J background and bred with *B6;129S6-Gt(ROSA)26Sor*$^{tm14(CAG\text{-}tdTomato)/Hze}$*/J* (Ai14; JAX 007908) mice, as previously described[38]. Both male and female mice were used for TRAP2 labelling experiments. Mice were injected retro-orbitally with rAAV9-mTNT-ChRmine-oScarlet or a vehicle control at three to four weeks of age. Four weeks later, mice were handled and acclimatized to fresh clean cages and optical pacing equipment for at least seven days before labelling. On the day of labelling, mice were allowed to acclimatize to optical pacing equipment for at least 2 h in a fresh clean cage with food and water, stimulated for 15 min (10-ms pulse width at 15 Hz for 500 ms every 1,500 ms) and left undisturbed for 2 h, at which time they were injected intraperitoneally with 5 mg kg^{-1} 4-hydroxytamoxifen (Sigma) dissolved in normal saline containing 1% Tween-80 and 2.5% DMSO (as described previously[37,39]). Mice were then returned to their home cage and were euthanized at least two weeks later to allow for full expression of the fluorophore.

Whole-brain CLARITY and analysis

Mice were perfused with ice-cold phosphate-buffered saline (PBS) and 4% paraformaldehyde (PFA), then post-fixed in a 1% CLARITY hydrogel solution (1% acrylamide, 0.003125% bis-acrylamide, 4% PFA and 0.25% VA-044 in 1× PBS) for 2 days. Tissue was degassed, polymerized at 37 °C for 4 h and washed with 200 mM sodium borate with 4% sodium dodecyl sulfate solution overnight. Tissue was then electrophoretically cleared for 3–7 days at 80 V (Life Canvas), passively cleared for an additional 2 days, then washed in PBS containing 0.2% Triton-X and 0.02% sodium azide at least 6 times at 37 °C. Cleared samples were refractive-index-matched using RapiClear (Sunjin Labs) and imaged on a custom-built light-sheet microscope[62] using a 10× objective and 5-µm step size or an LaVision Ultramicroscope with a 0.63× zoom macro lens with a step size of 5 µm. Images were visualized using Vision4D (Arivis).

For automated whole-brain registration and cell-segmentation analysis, images were loaded onto Arivis Vision4D software, and neurons were segmented using a built-in supervised pixel-based classifier package based on Ilastik[63] ('Trainable Segmenter'). Segmentation masks were converted to binary cell masks. Raw light-sheet microscopes images and cell masks were registered to a common reference space defined by the Allen Institute's Reference Atlas and analysed in a region-based manner using our MIRACL package[40].

Induction of *Fos* after pacing

Mice were injected retro-orbitally with AAV9-mTNT-ChRmine-oScarlet or vehicle at three to four weeks of age. At four weeks after injection, mice were handled and acclimatized to fresh clean cages and optical-pacing equipment for a minimum of seven days before pacing experiments. On the day of labelling, mice were allowed to acclimatize to optical-pacing equipment for at least 2 h in a fresh clean cage with food and water, stimulated for 15 min and euthanized 30 min after stimulation by perfusion with ice-cold PBS and 4% PFA under heavy anaesthesia. Tissue was post-fixed in 4% PFA on ice for an additional 24 h (brain) before staining and imaging.

In situ hybridization

Post-fixed brains were cut with a vibratome into 65-µm coronal slices. Heart and other organs were sliced at 200-µm thickness. Tissue slices were stored in 70% ethanol at −20 °C. Established protocols for third-generation hairpin chain reaction (HCR) in situ hybridization were used for coronal slice[64]. In situ hybridization probes (ChRmine, *Fos* and *Slc6a2*) were designed by and purchased from Molecular Instruments. Hybridization was performed overnight in hybridization buffer (Molecular Instruments) at 4 nM probe concentration. The next day, slices were washed (three times in wash buffer at 37 °C then twice in 2× SSCT at room temperature; 30 min each) and then incubated in amplification buffer. Dye-conjugated hairpins (B1-647, B3-488 and B5-546) were heated to 95 °C for 1 min and then cooled to 4 °C. Hairpin amplification was performed by incubating individual slices in 50 µl of amplification buffer with B1, B3 and B5 probes at concentrations of 240 nM overnight in the dark. Samples were stained with DAPI, washed three times with 5× SSCT for 30 min each and then equilibrated in exPROTOS (125 g iohexol, 3 g diatrizoic acid and 5 g *N*-methyl-D-glucamine dissolved in 100 ml deionized water with the refractive index adjusted to 1.458) (ref. [65]), a high-refractive-index mounting solution, then imaged. Slices were imaged on a confocal microscope (Olympus FV3000).

Cardiac histology

At 48 h post-fixation, hearts were sectioned into 200-µm slices. For staining, slices were first incubated for 10 min in blocking solution (3% normal donkey serum (NDS) in PBST), followed by primary antibody staining overnight at 4 °C using the following antibodies: anti-vimentin (ab24525), anti-cardiac troponin I (ab188877) or anti-PGP9.5 (ab108986), purchased from Abcam at 1:200 dilution in blocking solution. Slices were then washed twice in PBST, then stained with secondary antibodies (1 mg ml^{-1}) at 1:500 dilution for 3 h at room temperature using the following: F(ab')2 anti-chicken 488 (703-546-155) and anti-rabbit 647 (711-606-152) purchased from Jackson ImmunoResearch Laboratories. The slices were then stained with DAPI and washed three times with PBST (30 min per wash). Sections were mounted onto slides and mounted with exPROTOS. Slices were imaged on a confocal microscope (Olympus FV3000).

Stereotaxic surgery for optogenetic experiments

For all surgeries, mice were anesthetized with 1–2% isoflurane, and placed in a stereotaxic apparatus (Kopf Instruments) on a heating pad (Harvard Apparatus). Fur was removed from the scalp, the incision site was cleaned with betadine and a midline incision was made. Sterile surgical techniques were used, and mice were injected with sustained-release buprenorphine for post-operative recovery. Mice were allowed to recover for at least two weeks after surgery before behavioural experiments.

For intracranial optogenetic experiments, virus was injected using a 33-gauge beveled needle and a 10-μl Nano-fil syringe (World Precision Instruments), controlled by an injection pump (Harvard Apparatus). Five hundred nanolitres of AAVdj-hSyn::iC++-eYFP or AAVdj-hSyn::eYFP (5×10^{11} vg ml^{-1}) was injected at 150 nl per min and the syringe was left in place for at least 10 min before removal. The following coordinates were used (relative to Bregma): posterior insula (−0.58 (anterior–posterior (AP)), ±4.2 (medial–lateral (ML)), −3.85 (dorsal–ventral (DV)); mPFC (1.8 (AP), ±0.35 (ML), −2.9 (DV)). Optical fibres (0.39 NA, 200 μm; Thorlabs) were implanted 200 μm above virus injection coordinates. Fibres were secured to the cranium using Metabond (Parkell). Mice were allowed to recover for at least two weeks before behavioural testing.

In vivo electrophysiology

The mice with or without cardiac-targeted ChRmine expression were implanted with custom-made headplates, reference electrodes and cyanoacrylate-adhesive-based 'clear-skull caps' as previously described[66]. After recovery, mice were water-restricted and habituated to head fixation, but they were allowed to drink water to satiate thirst before recording sessions. Craniotomies were made with a dental drill at least several hours before recording sessions and were sealed with Kwik-Cast (World Precision Instruments). Exposed craniotomies before, during and after recordings were kept moist with frequent application of saline until sealed with Kwik-Cast.

Before recordings, the mice were placed into the pacemaker vests and reliable pacing was confirmed by ECG under brief anaesthesia with isoflurane. Then the mice were head-fixed and allowed to recover. Next, one or two (for simultaneous bilateral recordings) four-shank Neuropixels 2.0 probes mounted on a multi-probe manipulator system (New Scale Technologies) and controlled by SpikeGLX software (Janelia Research Campus) were inserted through the craniotomies at variable angles (0–20°) depending on the recording geometry. Typically the probes were aimed to touch the skull around the insula, which could be inferred from probe bending or changes in local field potential, and then were retracted around 100 μm and allowed to sit in place for at least 15 min before recordings. Recordings were performed along each of the four shanks sequentially while mice received 5 s of optical stimulation (900 bpm (15 Hz)) with inter-trial intervals of at least 15–25 s. Probes were cleaned with trypsin between recording sessions. Spike sorting was performed by Kilosort 2.5 and auxiliary software as previously described[66].

After recordings, the brains were perfused, cleared, imaged and registered to the Allen Brain Atlas as previously described[66]. Using the traces of lipophilic dye CM-DiI or DiD (which coated the probes before each insertion) and electrophysiological features, the atlas coordinates of the recorded single units were determined.

The spikes from single units were aligned to pacing onset, and the visualized peri-stimulus time histograms were calculated by subtracting 5 s baseline firing rate, 10 ms binning and 500 ms half-Gaussian filtering. The population-averaged firing rate of each region was calculated by combining z-scores (before filtering) over time for all single units in the region of interest. Specifically, we used hierarchical bootstrap to combine data from multiple levels as previously described[66]. For each condition, 100 bootstrap datasets were generated, and their mean and s.d. represented the mean and s.e.m. of the initial dataset. For statistical tests comparing ChRmine and control groups, the one-sided P value for the null hypothesis (the ChRmine firing rate subtracted by the control firing rate is zero) was calculated as the fraction of these subtracted values from the pairs of the resampled means (averaged over the time window of interest) that were smaller than zero.

Optogenetic freely moving behaviour

For optogenetic inhibition of iC++, a 473-nm laser (Omicron Laserage) was used to deliver constant light at 2–3 mW measured from the tip. Laser shutters were controlled using a Master-8 receiving synchronized input from behaviour apparatus and control software (Ethovision).

Statistical analysis

The target number of subjects used in each experiment was determined on the basis of numbers in previously published studies. No statistical methods were used to predetermine sample size or randomize. Criteria for excluding mice from analysis are listed in the methods. Mean ± s.e.m. was used to report statistics. The statistical test used, definition of n and multiple-hypothesis correction where appropriate are described in the figure legends. Unless otherwise stated, all statistical tests were two-sided. Significance was defined as alpha = 0.05. All statistical analyses were performed in GraphPad Prism 9.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All primary data for all figures and extended data figures are available from the corresponding author upon request.

62. Tomer, R., Ye, L., Hsueh, B. & Deisseroth, K. Advanced CLARITY for rapid and high-resolution imaging of intact tissues. *Nat. Protoc.* **9**, 1682–1697 (2014).
63. Berg, S. et al. ilastik: interactive machine learning for (bio)image analysis. *Nat. Methods* **16**, 1226–1232 (2019).
64. Choi, H. M. et al. Third-generation in situ hybridization chain reaction: multiplexed, quantitative, sensitive, versatile, robust. *Development* **145**, dev165753 (2018).
65. Park, Y.-G. et al. Protection of tissue physicochemical properties using polyfunctional crosslinkers. *Nat. Biotechnol.* **37**, 73 (2018).
66. Sylwestrak, E. L. et al. Cell-type-specific population dynamics of diverse reward computations. *Cell* **185**, 3568–3587 (2022).

Acknowledgements We thank all members of the K.D. laboratory, particularly M. Lovett-Barron, C. Bedbrook and F. Gore, for comments. This work was supported by grants from the NIH, NSF, Gatsby, Fresenius, Wiegers, Grosfeld and NOMIS Foundations (to K.D.); the Stanford MSTP and Bio-X (to B.H.); and a K99 NS119784 and NARSAD Young Investigator Grant (to R.C.).

Author contributions B.H., R.C. and K.D. designed the project and experiments, and wrote the manuscript. B.H., R.C., D.T., M.R., Y.S.K., M.I., S.R., S.P., D.K.K., S.H.K., L.T., L.M., A. Cordero, J.S., M. Zhao, T.H., A. Crow, A.-C.W.Y., C. Raja, K.E. and D.B. contributed to the collection of data and interpretation of results with input from K.D. Y.J. performed electrophysiology recordings and analysis with assistance from T.X.L. and R.C. C. Ramakrishnan designed the viral constructs. B.H., M.G. and M. Zeineh performed whole-brain image registration. K.D. supervised all aspects of this work.

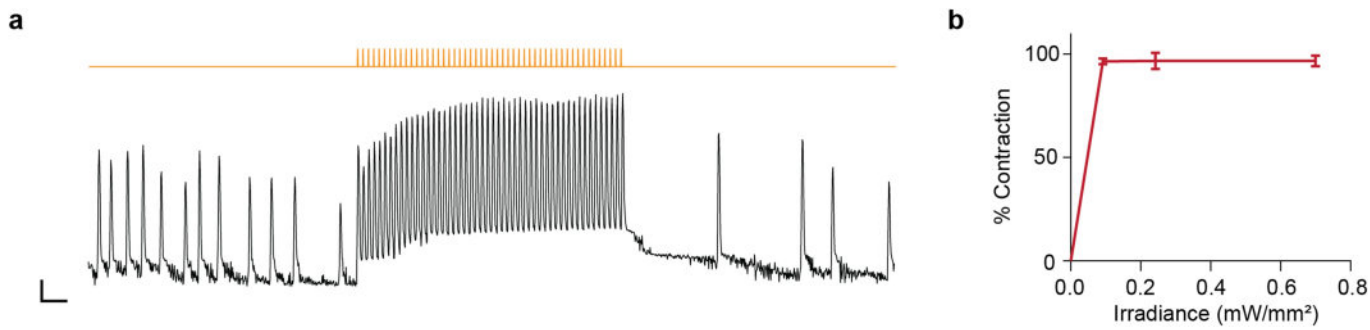
Competing interests The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-05748-8.

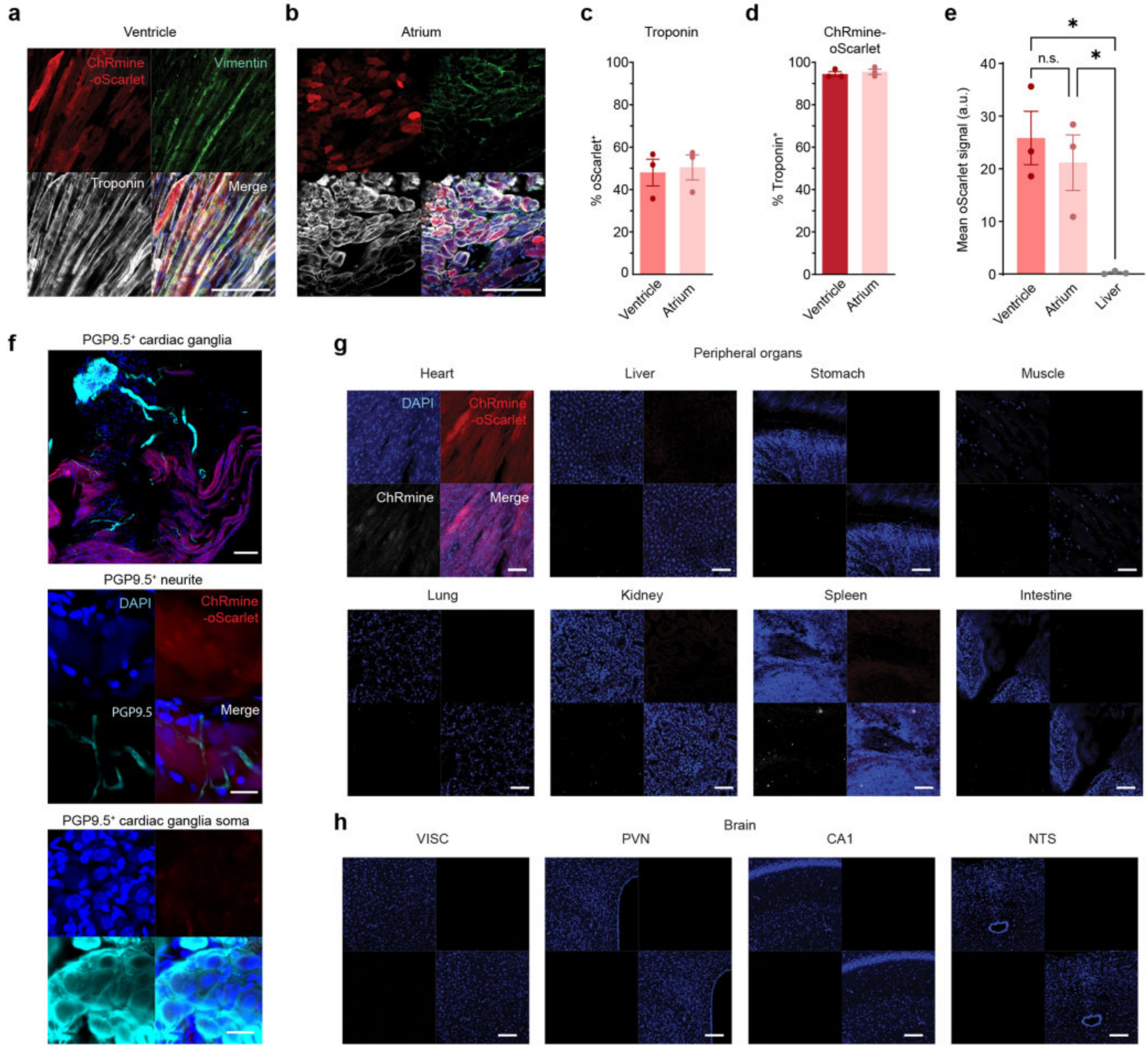
Correspondence and requests for materials should be addressed to Karl Deisseroth.

Peer review information *Nature* thanks Anna Beyeler, Tobias Bruegmann, Emilia Entcheva and Scott Russo for their contribution to the peer review of this work.

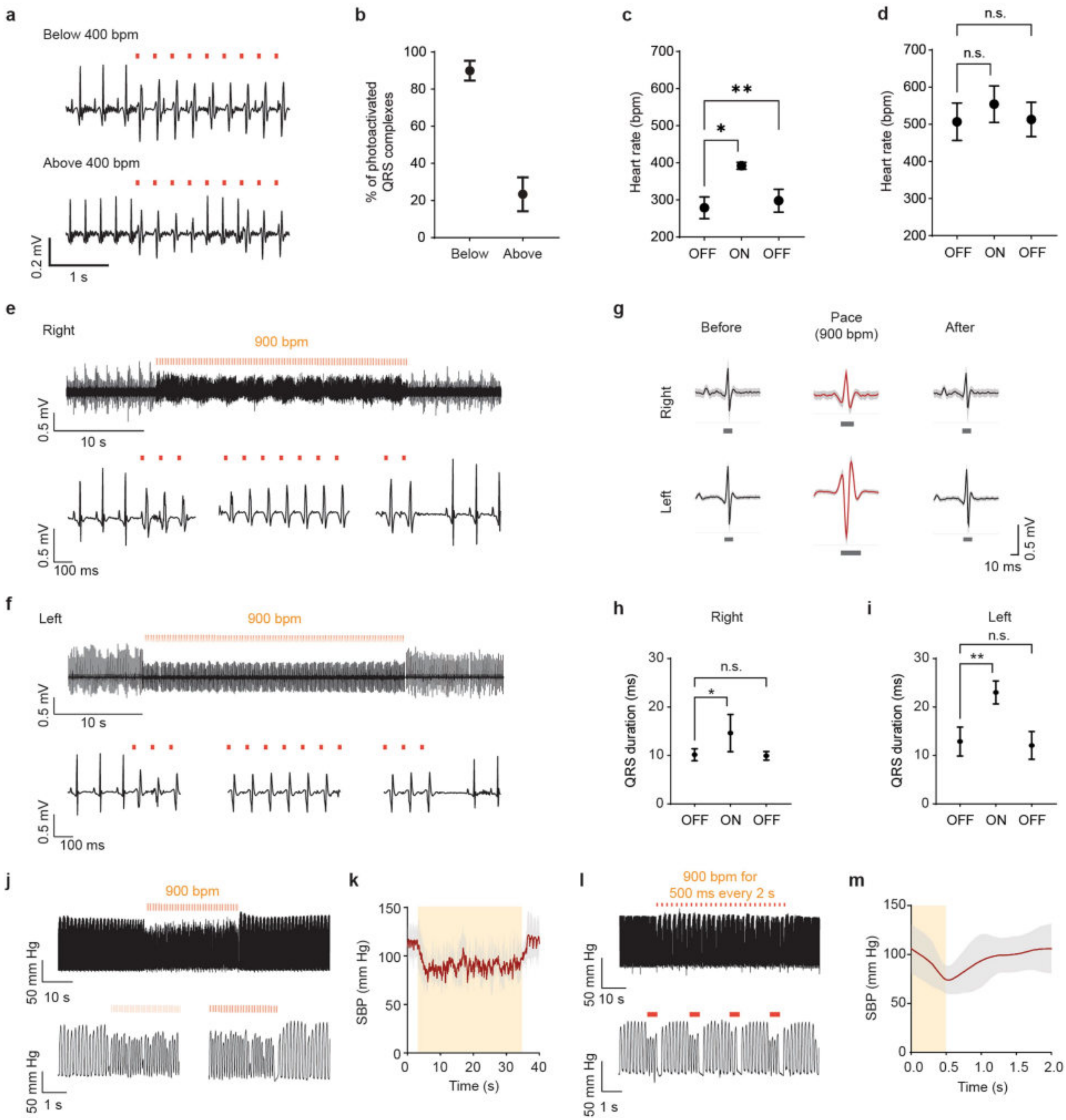
Reprints and permissions information is available at http://www.nature.com/reprints.



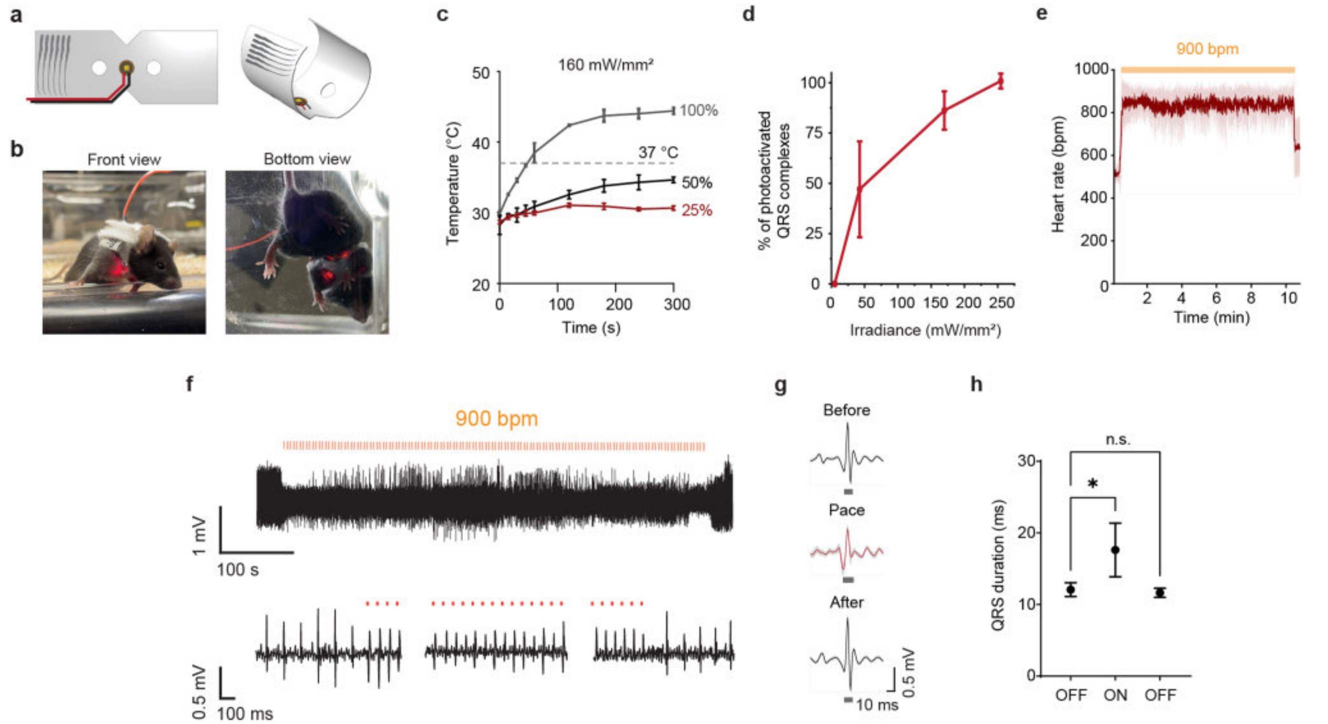
Extended Data Fig. 1 | In vitro characterization of the optical pacemaker. **a**, Quantified cardiomyocyte contraction sequence measured by centroid motion (see Supplementary Video 1). Optical stimulation was applied at 5 Hz, with 10-ms pulse width at 585 nm. Scale: 1 s, 1 pixel. **b**, Fidelity of optically induced contractions at different light intensities. n = 20 cells. Data represent mean ± s.e.m.



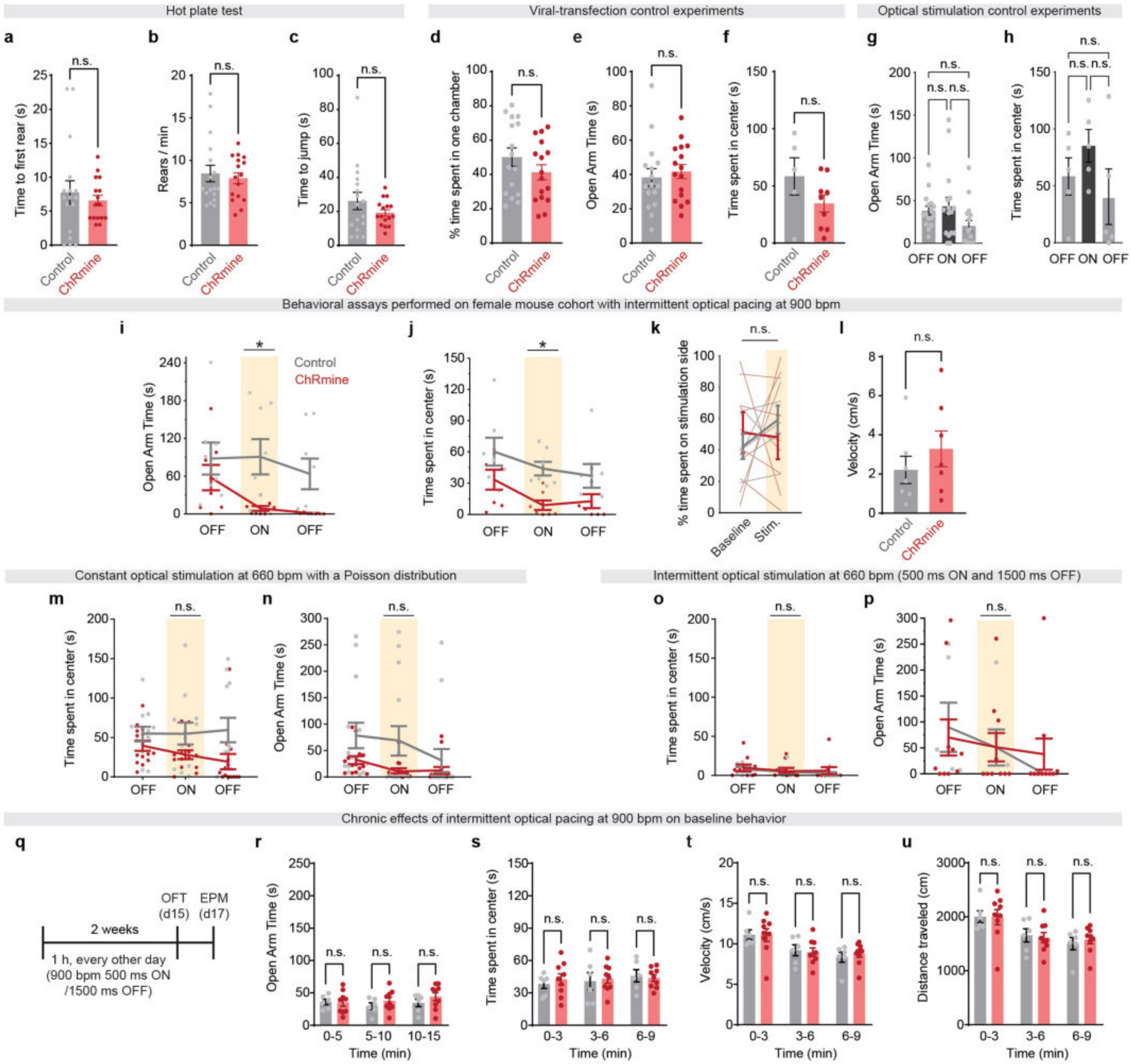
Extended Data Fig. 2 | In vivo characterization of AAV9-mTNT::ChRmine-2A-oScarlet. a,**b**, Representative confocal images of ChRmine-2A-oScarlet (red)-infected ventricular (**a**) and atrial (**b**) cardiac tissue, co-labelled with troponin (white), vimentin (green), and DAPI (blue). Note ChRmine expression is restricted to troponin+ cardiomyocytes with no off-target labelling in neighbouring vimentin+ fibroblasts. Scale bar=100 µm. **c**, Penetrance of AAV9-mTNT::ChRmine-p2A-oScarlet quantified as percentage of troponin+ cells that express ChRmine-2A-oScarlet (n = 3). **d**, Specificity of AAV9-mTNT::ChRmine-p2A-oScarlet quantified as percentage of ChRmine-2A-oScarlet+ cells that are troponin+ (n = 3). **e**, Quantification of oScarlet expression in ventricular and atrial cardiac tissue and liver as mean fluorescence signal (arbitrary unit) following 1 month post-injection of AAV9-mTNT::ChRmine-p2A-oScarlet (n = 3, one-way repeated-measures ANOVA with Bonferroni post-hoc test: $F(2,4)=18.3$, $p = 0.0097$. Post-hoc: ventricle vs. atrium $p = 0.99$, ventricle vs. liver $p = 0.014$, atrium vs liver $p = 0.029$). **f**, Representative confocal image from one of two mice depicting lack of neuronal labelling in cardiac ganglia following retro-orbital delivery of AAV9-mTNT::ChRmine-p2A-oScarlet as measured by immunostaining for the neural marker PGP9.5 (cyan). No off-target ChRmine expression was observed in neurites (centre) or in the soma of cardiac ganglia (bottom). Scale bar=100 µm (top), 20 µm (center, bottom). **g**,**h**, Representative confocal images of oScarlet (red), ChRmine mRNA (white) and nuclei (DAPI) nuclei across different organs (**g**) after 9 months of expression from one of two mice. No off-target expression of ChRmine was observed in organs beyond the heart, including throughout the brain (**h**). Scale bar=100 µm. Data represent mean ± s.e.m.



Extended Data Fig. 3 | Quantification of cardiac responses to optical pacing in vivo. a, Representative ECG recording with optical pacing at 400 bpm below (top) or above (bottom) a resting heart rate of 400 bpm. **b**, Percentage of photoactivated heart beats that track with the delivered 400 bpm optical stimulus below or above a resting heart of 400 bpm (n = 4 (below), n = 5 (above) mice). **c**, Measured heart rate before, during, and after pacing at 400 bpm in heavily anesthetized mice with a resting heart rate below 400 bpm (n = 4 mice, one-way repeated-measures ANOVA with Bonferroni post-hoc test: condition $F_{(1.01, 3.03)}$=34.07, p = 0.0097; individual $F_{(3,6)}$=2.4, p = 0.17. Post-hoc: OFF vs. ON *p = 0.015, OFF vs. OFF **p = 0.002). **d**, Measured heart rate before, during, and after pacing at 400 bpm in lightly anesthetized mice with a resting heart rate above 400 bpm (n = 5 mice, one-way repeated-measures ANOVA with Bonferroni post-hoc test: condition $F_{(1.06, 4.24)}$=3.71, p = 0.12; individual $F_{(4,8)}$=5.9, p = 0.016. Post-hoc: OFF vs. ON p = 0.17, OFF vs. OFF p = 0.72). **e**, Representative ECG recording with optical pacing delivered at 900 bpm with 10-ms pulse width with laser positioned at the centre of the chest to induce right ventricular pacing (top) with additional traces from before, during, and after light stimulation (bottom). **f**, Representative ECG recording with optical pacing delivered at 900 bpm with 10-ms pulse width with laser positioned more lateral of the chest to induce left ventricular pacing (top) with additional traces from before, during, and after light stimulation (bottom). **g**, Representative QRS complexes averaged over 100 heart beats before, during, and after pacing for right and left ventricular stimulation. Dark grey bar indicates duration of QRS complex. **h**, QRS duration before, during, and after right ventricular pacing (n = 4 mice, one-way repeated-measures ANOVA with Bonferroni post-hoc test: condition $F_{(1.01, 3.04)}$=25.4, p = 0.015; individual $F_{(3,6)}$=0.68, p = 0.59. Post-hoc: OFF vs. ON *p = 0.037, OFF vs. OFF p = 0.99). **i**, QRS duration before, during, and after left ventricular pacing (n = 4 mice, one-way repeated-measures ANOVA with Bonferroni post-hoc test: condition $F_{(1.18, 3.54)}$=121.7, p = 6.7e-4; individual $F_{(3,6)}$=16.64, p = 0.0026. Post-hoc: OFF vs. ON **p = 0.0038, OFF vs. OFF p = 0.20). **j**, Representative left ventricular blood pressure recordings with sustained optical pacing delivered at 900 bpm with 10-ms pulse width (top) with additional traces from before and after light stimulation (bottom). **k**, Systolic blood pressure (SBP) over time with sustained 900 bpm optical pacing (orange) for 30 s showing a sustained drop in SBP during stimulation (n = 3 mice). **l**, Representative left ventricular blood pressure recordings with intermittent optical pacing delivered at 900 bpm with 10-ms pulse width for 500 ms every 1,500 ms (top) with additional traces during light stimulation (bottom). **m**, Averaged SBP over the interval of optical stimulation (500 ms ON at 900 bpm, 1,500 ms OFF) (n = 3 mice). Data represent mean ± s.d.

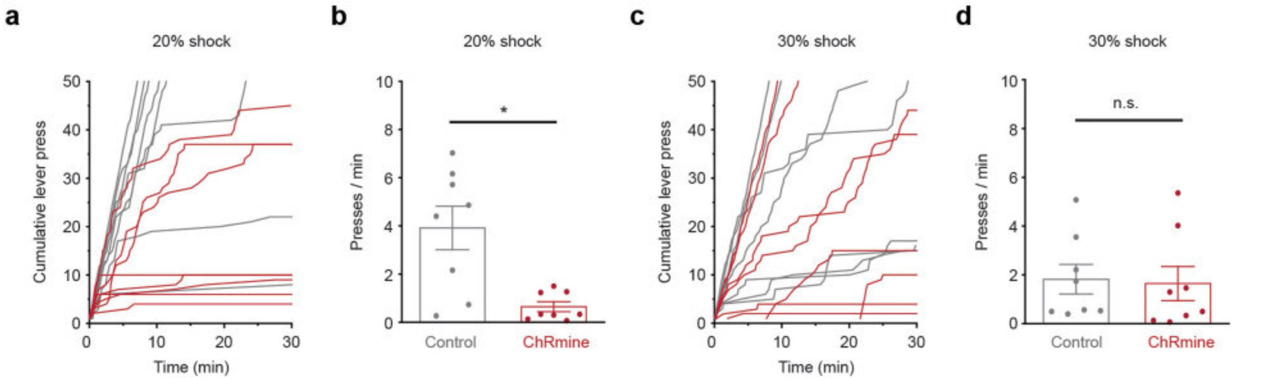


Extended Data Fig. 4 | Characterization of the wearable micro-LED vest. **a**, Schematic of optical pacing vest. **b**, Photographs of freely moving mouse wearing optical pacing vest while receiving 591 nm light stimulation. **c**, Heating of device operated at 900 bpm with 10-ms pulse width at varying duty cycles (n = 3 devices, 25% (500 ms ON, 1,500 ms OFF), 50% (1 s ON, 1 s OFF), 100% (constant ON)). **d**, Percentage of photoactivated QRS complexes as a function of irradiance using the micro-LED device (n = 5). **e**, Measured heart rate across time over a 10 min stimulation period (900 bpm) showing ability of optical pacing to induce a stable heart rhythm (n = 6 mice). **f**, Representative ECG recording from one mouse receiving 900-bpm stimulation for 10 min (top) with additional traces from before, during, and after light stimulation (bottom). **g**, Representative QRS complexes averaged over 100 heart beats before, during and after pacing. **h**, QRS duration before, during and after right ventricular pacing (n = 5 mice, one-way repeated-measures ANOVA with Bonferroni post-hoc test: condition $F_{(1.03, 4.10)}$=13.74, p = 0.02; individual $F_{(4,8)}$=1.83, p = 0.21. Post-hoc: OFF vs. ON *p = 0.024, OFF vs. OFF p = 0.88). Data represent mean ± s.e.m.

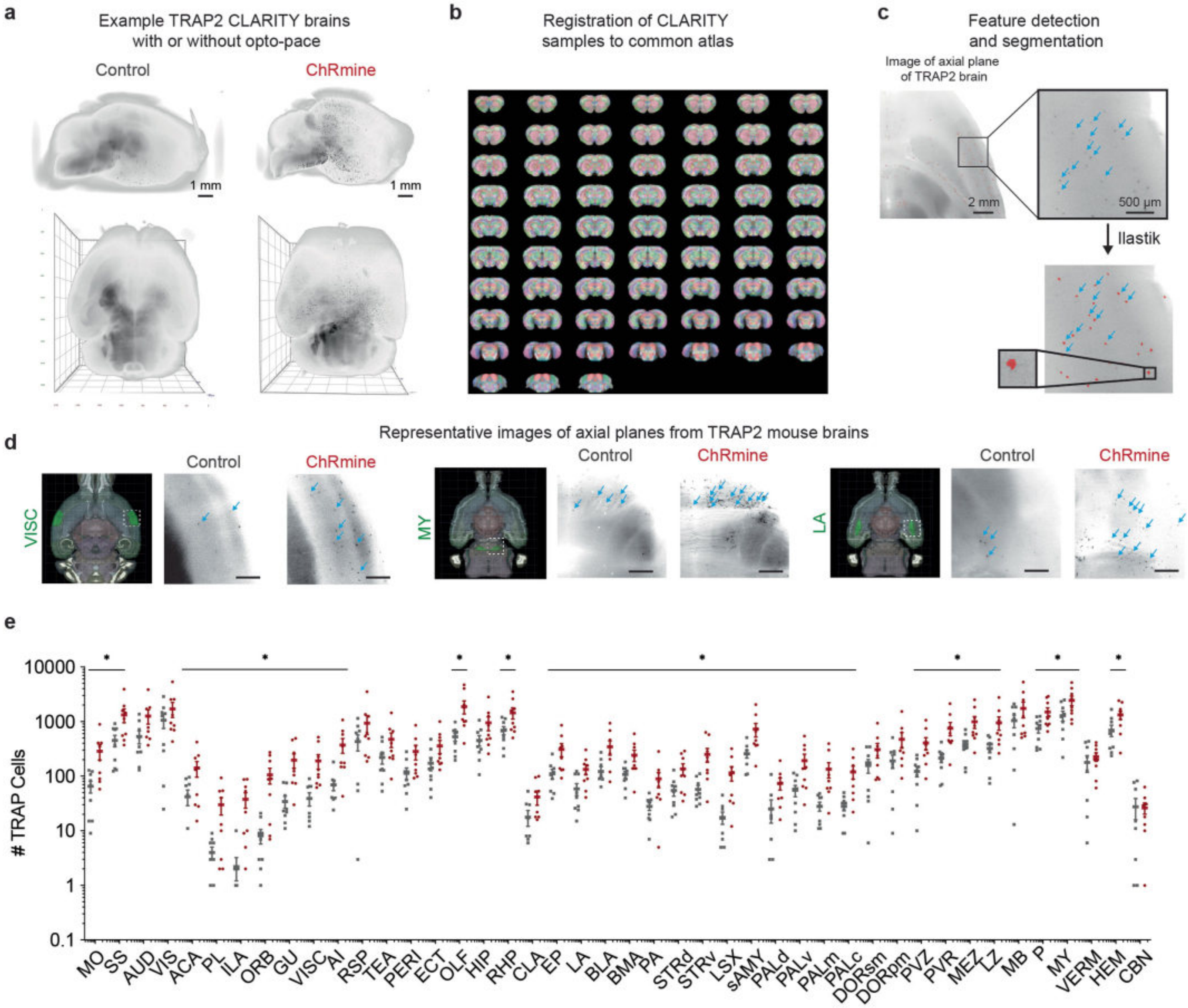


Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Additional characterization of optical-pacing effects on mouse behaviour. a–**c**, A hot-plate test was performed to assess for potential effects on thermal pain thresholds from optical pacing (n = 17 (control), 16 (ChRmine)) and the following were quantified: time to first rear (unpaired two-tailed t-test, p = 0.53) (**a**); rears per minute (unpaired two-tailed t-test, p = 0.64) (**b**); and time to first jump (unpaired two-tailed t-test, p = 0.22) (**c**). Note no statistical significance in thermal thresholds was observed with cardiac pacing. **d**–**f**, To assess behavioural differences between control and virally transduced ChRmine-expressing mice, the following comparisons were performed: time spent in one chamber during baseline day (no light delivery) (n = 16 per group, unpaired two-tailed t-test p = 0.21) (**d**); time spent in the open arm of the EPM test during the first 5-min epoch with no light delivery (n = 16 per group, unpaired two-tailed t-test p = 0.61) (**e**); and time spent in the centre of the OFT during the first 3-min epoch with no light delivery (n = 5 (control), 9 (ChRmine), unpaired two-tailed t-test p = 0.15) (**f**). Note no statistical significance in behaviour was observed from viral-transfection. **g**,**h**, To assess for effects of light stimulation alone on mouse behaviour, the following comparisons were performed: time spent in the open arms by control (saline injected) mice during the 15 min EPM assay, where intermittent light delivery (10-ms pulse width, 900 bpm for 500 ms every 2 s) was delivered during the 5 min ON epoch (n = 16, one-way repeated-measures ANOVA with Bonferroni post-hoc test: condition $F_{(1.44,21.62)}$=2.942, p = 0.088, individual $F_{(15,30)}$=1.61, p = 0.13. Post-hoc: OFF vs. ON p = 0.99, OFF vs. OFF p = 0.9, ON vs. OFF p = 0.27) (**g**); and time spent in the centre by control mice during the 9 min OFT, where intermittent light delivery was delivered during the 3 min ON epoch (n = 5, one-way repeated-measures ANOVA with Bonferroni post-hoc test: condition $F_{(1.84,7.45)}$=1.67, p = 0.25, individual $F_{(4,8)}$=1.51, p = 0.29. Post-hoc: OFF vs. ON p = 0.99, OFF vs. OFF p = 0.99, ON vs. OFF p = 0.41) (**h**). Note no statistical significance in behaviour was observed from light stimulation alone. **i**,**j**, To assess for effects from optical pacing (10-ms pulse width, 900 bpm for 500 ms every 2 s) on anxiety-like behaviour in female mice, the following behavioural assays were measured: time spent in the open arms during the EPM (n = 8 (control) and 7 (ChRmine) female mice, two-way repeated-measures ANOVA with Bonferroni post-hoc test: group (opsin) x time interaction $F_{(2,26)}$=1.91, p = 0.17, group (opsin) effect $F_{(1,13)}$=5.1, p = 0.042; time effect $F_{(2,26)}$=4.24, p = 0.026. Bonferroni post-hoc: ON epoch ChRmine vs Control *p = 0.033) (**i**); and time spent in the centre during the OFT (n = 7 (control) and 7 (ChRmine) female mice, two-way repeated-measures ANOVA with Bonferroni post-hoc test: group (opsin) x time interaction $F_{(2,24)}$=0.37, p = 0.70; group (opsin) effect $F_{(1,12)}$=7.47, p = 0.018; time effect $F_{(2,24)}$=6.9, p = 0.0043. Post-hoc: ON epoch ChRmine vs Control *p = 0.031) (**j**). **k**,**l**, To determine whether cardiac pacing was aversive in female mice, we also measured the percentage of time spent on baseline and stimulation day for control (grey) and ChRmine-expressing (red) mice during the RTPP assay (n = 7 female mice per group, two-way repeated-measures ANOVA with Bonferroni post-hoc test: group (opsin) x treatment interaction $F_{(1,12)}$=0.68, p = 0.42; group (opsin) effect $F_{(1,12)}$=0.11, p = 0.91; treatment effect $F_{(1,12)}$=0.35, p = 0.57. Post-hoc: Baseline vs Stimulation for Control: p = 0.67, ChRmine: p = 0.99) (**k**); and the average velocity on the optically paced side during RTPP (n = 7 female mice per group, two-way repeated-measures ANOVA with Bonferroni post-hoc test: group (opsin) x treatment interaction $F_{(1,12)}$=0.088, p = 0.77; group (opsin) effect $F_{(1,12)}$=3.69, p = 0.079; treatment effect $F_{(1,12)}$=0.12, p = 0.73. Post-hoc: Stimulation day Control vs Chrmine: p = 0.59) (**l**). **m**,**n**, To determine whether increased heart rate variability can affect anxiety-like behaviour, we introduced constant 660 bpm stimulation with a Poisson distribution in mice and measured the time spent in the centre during the OFT (n = 12 (control) and 14 (ChRmine) mice, two-way repeated-measures ANOVA with Bonferroni post-hoc test: group (opsin) x time interaction $F_{(2,48)}$=1.29, p = 0.28; group (opsin) effect $F_{(1,24)}$=5.80, p = 0.024; time effect $F_{(1.36,32.6)}$=0.47, p = 0.55. Post-hoc ChRmine vs Control: OFF p = 0.61, ON p = 0.28, OFF p = 0.12) (**m**); and the time spent in the open arms during the EPM (n = 14 (control) and 14 (ChRmine) mice, two-way repeated-measures ANOVA with Bonferroni post-hoc test: group (opsin) x time interaction $F_{(2,52)}$=1.75, p = 0.18; group (opsin) effect $F_{(1,26)}$=3.3, p = 0.082; time effect $F_{(1.84,47.8)}$=5.157, p = 0.011. Post-hoc ChRmine vs Control: OFF p = 0.25, ON p = 0.20, OFF p = 0.99) (**n**). **o**,**p**, To determine whether intermittent tachycardia at a rhythm below 900 bpm can affect anxiety-like behaviour, we introduced intermittent 660 bpm stimulation (10-ms pulse width, 660 bpm for 500 ms every 2 s) in mice and measured the time spent in the centre during the OFT (n = 6 (control) and 10 (ChRmine) mice, two-way repeated-measures ANOVA with Bonferroni post-hoc test: group (opsin) x time interaction $F_{(2,28)}$=0.13, p = 0.88; group (opsin) effect $F_{(1,14)}$=0.25, p = 0.63; time effect $F_{(2,28)}$=1.1, p = 0.35. Post-hoc ChRmine vs Control: OFF p = 0.99, ON p = 0.99, OFF p = 0.99) (**o**); and the time spent in the open arms during the EPM (n = 6 (control) and 10 (ChRmine) mice, two-way repeated-measures ANOVA with Bonferroni post-hoc test: group (opsin) x time interaction $F_{(2,28)}$=0.52, p = 0.60; group (opsin) effect $F_{(1,14)}$=0.03, p = 0.86; time effect $F_{(2,28)}$=2.23, p = 0.12. Post-hoc ChRmine vs Control: OFF p = 0.99, ON p = 0.99, OFF p = 0.99) (**p**). **q**, Schematic overview of the chronic stimulation experiment. Mice were stimulated with intermittent optical pacing (900 bpm for 500 ms every 2 s) for 1 h every other day for two weeks before performing the OFT and EPM behavioural assays. **r**, Time spent in open arm during the EPM test for control (grey) and ChRmine (red) mice subjected to chronic optical stimulation (n = 6 (control) and 9 (ChRmine) mice, two-way repeated-measures ANOVA with Bonferroni post-hoc test: group (opsin) x time interaction $F_{(2,26)}$=0.87, p = 0.43; group (opsin) effect $F_{(1,13)}$=0.71, p = 0.41; time effect $F_{(1.65,21.4)}$=0.96, p = 0.38. Post-hoc ChRmine vs Control: 0–5 min p = 0.99, 5–10 min p = 0.89, 10–15 min p = 0.84). **s**, Time spent in centre during the OFT test for control (grey) and ChRmine (red) mice subjected to chronic optical stimulation (n = 6 (control) and 9 (ChRmine) mice, two-way repeated-measures ANOVA with Bonferroni post-hoc test: group (opsin) x time interaction $F_{(2,26)}$=0.25, p = 0.78; group (opsin) effect $F_{(1,13)}$=0.097, p = 0.76; time effect $F_{(1.78,23.2)}$=0.49, p = 0.60. Post-hoc ChRmine vs Control: 0–3 min p = 0.99, 3–6 min p = 0.99, 6–9 min p = 0.99). **t**, Average velocity (cm/s) of mice in the OFT test for control (grey) and ChRmine (red) mice subjected to chronic optical stimulation (n = 6 (control) and 9 (ChRmine) mice, two-way repeated-measures ANOVA with Bonferroni post-hoc test: group (opsin) x time interaction $F_{(2,26)}$=0.29, p = 0.78; group (opsin) effect $F_{(1,13)}$=0.005, p = 0.94; time effect $F_{(1.34,17.4)}$=17.3, p = 2.6e-4. Post-hoc ChRmine vs Control: 0–3 min p = 0.99, 3–6 min p = 0.99, 6–9 min p = 0.99). **u**, Total distance travelled (cm) mice during the OFT test for control (grey) and ChRmine (red) mice subjected to chronic optical stimulation (n = 6 (control) and 9 (ChRmine) mice, two-way repeated-measures ANOVA with Bonferroni post-hoc test: group (opsin) x time interaction $F_{(2,26)}$=0.29, p = 0.75; group (opsin) effect $F_{(1,13)}$=0.005, p = 0.94; time effect $F_{(1.34,17.4)}$=17.3, p = 2.6e-4. Post-hoc ChRmine vs Control: 0–3 min p = 0.99, 3–6 min p = 0.99, 6–9 min p = 0.99). Data represent mean ± s.e.m.



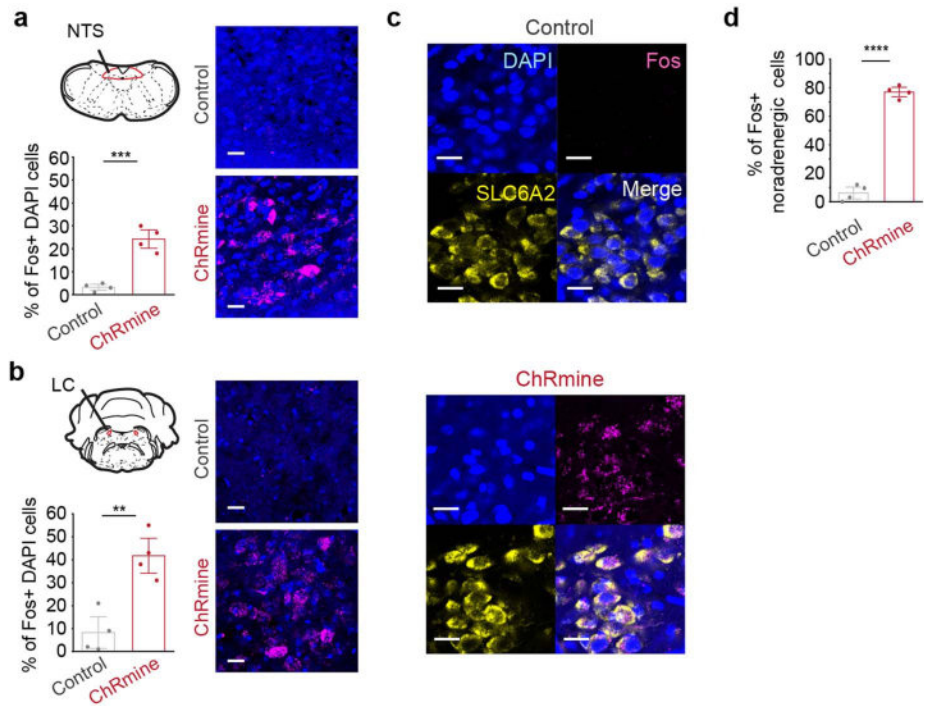
Extended Data Fig. 6 | Operant lever-pressing task at higher risk of shock. **a**, Cumulative lever presses during 20% shock session for control (grey) and ChRmine-expressing (red) mice (n = 8 mice per group). Experiment was performed otherwise identically to the behaviour described in Fig. 2. **b**, Lever-pressing rate averaged across the entire 20% shock trial session (n = 8 mice per group, Two-tailed *t*-test, *p = 0.0367). **c**, Cumulative lever presses during 30% shock session. **d**, Lever-pressing rate averaged across the entire 30% shock trial session (n = 8 mice per group, Two-tailed t-test, p = 0.7663). Data represent mean ± s.e.m.



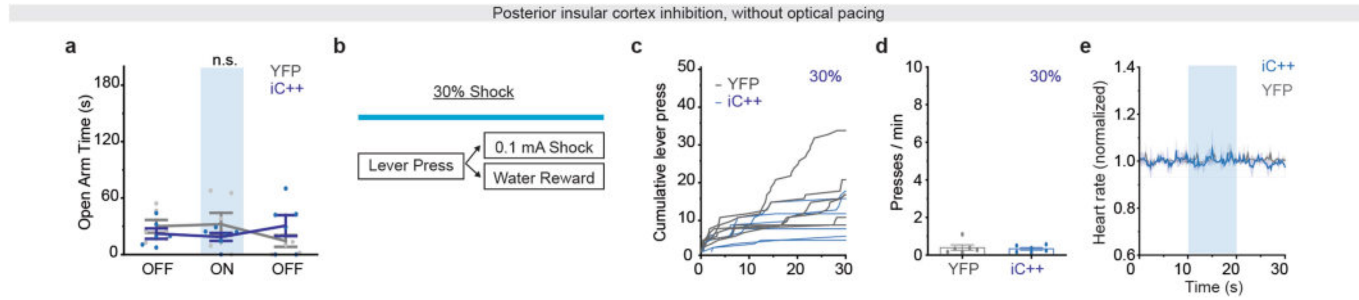
Acronym Table

MO	Somatomotor area	AI	Agranular insular area	LA	Lateral amygdalar nucleus	PALm	Pallidum, medial region	MY	Medulla
SS	Somatosensory area	RSP	Retrosplenial area	BLA	Basolateral amygdalar nucleus	PALc	Pallidum, caudal region	VERM	Cerebellar cortex, vermal regions
GU	Gustatory area	TEA	Temporal association area	BMA	Basomedial amygdalar nucleus	DORsm	Thalamus, sensory-motor cortex related	HEM	Cerebellar cortex, hemispheric regions
VISC	Visceral area	PERI	Perirhinal area	PA	Posterior amygdalar nucleus	DORpm	Thalamus, polymodal association cortex related	CBN	Cerebellar nuclei
AUD	Auditory areas	ECT	Ectorhinal area	STRd	Striatum dorsal region	PVZ	Hypothalamus, periventricular zone		
VIS	Visual areas	OLF	Olfactory area	STRv	Striatum ventral region	PVR	Hypothalamus, periventricular region		
ACA	Anterior cingulate area	HIP	Hippocampal region	LSX	Lateral septal complex	MEZ	Hypothalamic medial zone		
PL	Prelimbic area	RHP	Retrohippocampal region	sAMY	Striatum-like amygdalar nuclei	LZ	Hypothalamic lateral zone		
ILA	Infralimbic area	CLA	Claustrum	PALd	Pallidum, dorsal region	MB	Midbrain		
ORB	Orbital area	EP	Endopiriform nucleus	PALv	Pallidum, ventral region	P	Pons		

Extended Data Fig. 7 | Generation of brain-wide activity maps during optical pacing. a, Representative whole-brain CLARITY sagittal (top) and axial (bottom) images of control and ChRmine expressing mice exposed to optical pacing. Heart-targeted ChRmine expression resulted in increased tdTomato+ cells throughout the brain (seen as black dots) relative to control. **b**, Following light-sheet imaging, each image stack was registered to a common Allen Reference Atlas using our previously reported computational pipeline[40]. Depicted are coronal slices across the brain with the overlaid anatomical atlas, where different anatomical regions are depicted with different colours. **c**, Representative raw image of single plane of TRAP2-tdTomato brain imaged under light-sheet microscope (left), and after detection by a supervised classifier (Ilastik, Arivis plugin), where single-cells are outlined in red (right, bottom). Blue arrows indicate example features that were detected by the classifier. **d**, Example CLARITY light-sheet images of control and ChRmine-paced mice in the visceral cortex (VISC), Medulla-behavioural state related (MY), and in the lateral amygdalar nucleus (LA), which are outlined in green in the axial reference slice. Scale bar=500 μm. **e**, Regional cell counts of paced (ChRmine, red) vs control (grey) cohorts sorted from anterior to posterior in all anatomical regions (n = 9 per group, multiple two-sided t-tests corrected for multiple comparisons with the Benjamini and Hochberg method (*FDR=10%). p-values: MO p = 0.025, SS p = 0.038, AUD p = 0.082, VIS p = 0.31, ACA p = 0.050, PL p = 0.027, ILA p = 0.0086, ORB p = 0.012, GU p = 0.013, VISC p = 0.018, AI p = 0.016, RSP p = 0.20, TEA p = 0.11, PERI p = 0.074, ECT p = 0.10, OLF p = 0.017, HIP p = 0.10, RHP p = 0.052, CLA p = 0.092, EP p = 0.035, LA p = 0.028, BLA p = 0.051, BMA p = 0.036, PA p = 0.043, STRd p = 0.029, STRv p = 0.015, LSX p = 0.012, sAMY p = 0.037, PALd p = 0.036, PALv p = 0.033, PALm p = 0.024, PALc p = 0.012, DORsm p = 0.16, DORpm p = 0.095, PVZ p = 0.016, PVR p = 0.020, MEZ p = 0.029, LZ p = 0.033, MB p = 0.28, P p = 0.030, MY p = 0.054, VERM p = 0.50, HEM p = 0.043, CBN p = 0.92). Acronyms for each major brain region is listed.



Extended Data Fig. 8 | Optically induced tachycardia increases *Fos* mRNA expression in brain regions associated with the central autonomic network. **a**,**b**, Percentage of cells that are *Fos*+ determined from in situ hybridization for *Fos* mRNA (magenta) following cardiac pacing in control and ChRmine-expressing mice in the nucleus of the solitary tract (NTS) (**a**) and in the locus coeruleus (LC) (**b**) (n = 4 mice per group, unpaired two-tailed t-test, ***p = 0.00029 (NTS), **p = 0.0027). We have also performed in situ hybridization for *Fos* mRNA in the nodose ganglion and observed potential induction in these vagal sensory neurons by the cardiac signals (n = 4, control 0.96% ± 0.7; ChRmine 5.29% ± 1.6, unpaired two-tailed t-test, p = 0.04). **c**, Confocal images of the locus coeruleus stained for *Fos* (magenta), the norepinephrine transporter *Slc6a2* (yellow) and DAPI (blue). **d**, Quantification of neurons that express *Slc6a2* that also co-localize with *Fos* during optical pacing (n = 4 mice per group, unpaired two-tailed t-test: ****p = 1.0e-7). Scale bar=20 µm. Data represent mean ± s.e.m.



Extended Data Fig. 9 | Behavioural and physiological effects from inhibition of the posterior insula. a, Mice were evaluated on the EPM with iC++ inhibition only during the "ON" epoch, with no optical pacing vest (n = 6 mice per group; two-way repeated-measures ANOVA: group (opsin) x time interaction $F_{(2,20)}$=2.569, p = 0.1016; group (opsin) effect $F_{(1,10)}$=0.0392, p = 0.8470; time effect $F_{(2,20)}$=0.1486, p = 0.8629). Note inhibition of pIC did not significantly alter time spent in the open arms of the EPM. **b**, Schematic overview of the pIC inhibition experiment. Mice were evaluated on the Vogel conflict task with elevated chance of shock (30%) while wearing the optical pacing vest but without receiving cardiac stimulation. Optogenetic inhibition was performed during the entire duration of the 30 min trial. **c**, Individual traces of lever presses during shock day with 30% chance of shock and bilateral pIC inhibition (n = 6 mice per group). **d**, Lever-pressing rate averaged across entire 30% shock session (n = 6 mice per group). **e**, Heart rate as a function of time during pIC inhibition (blue) (n = 3 mice per group). Data represent mean ± s.e.m.

nature portfolio

Corresponding author(s): Karl Deisseroth

Last updated by author(s): Dec 19, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Olympus Fluoview FV3000 (confocal image acquisition), Ethovision XT14 (behavioral tracking), Open Ephys (extracellular in vivo recording acquisition), MATLAB R2021b, SpikeGLX (neuropixels electrophysiology), Rodent Surgical Monitor+ Indus Instruments (Electrocardiography)
Data analysis	MATLAB R2021b, Python, ImageJ 1.53, Origin 2020, GraphPad Prism, Kilosort2.5, Phy2, Arivis Vision4D, MIRACL package for whole-brain registration, Ilastik for cell segmentation, Vision4D (Arivis)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All primary data for all figures and supplementary figures are available from the corresponding author upon request.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to determine sample size a priori. We chose sample sizes there were similar to those reported in our previous publications.
Data exclusions	Criterias for animal exclusion were pre-established. Animals with unintended experimental error were excluded based on poor viral expression from mis-injection or misplacement of fiber optics. Otherwise, no data were excluded.
Replication	Data was collected in multiple experiments (at least twice) and all results were pooled. No experiments failed to replicate. Behavior that required repeated measure were counter balanced.
Randomization	Mice were randomly assigned to each experimental group.
Blinding	Experimenters were not blinded to opsin identity during behavioral experiments because animals were run across several behavioral tasks, some of which required long training periods. Furthermore, electrophysiology recordings were performed in a difficult-to-target region, so opsin identity needed to be known in order to guarantee the best recording outcome. Experimenters were blinded to treatment conditions for in situ Fos and other histology quantification.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☐	☒	Antibodies
☒	☐	Eukaryotic cell lines
☒	☐	Palaeontology and archaeology
☐	☒	Animals and other organisms
☒	☐	Clinical data
☒	☐	Dual use research of concern

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging

Antibodies

Antibodies used	Abcam: Anti-Vimentin (ab24525, 1:200), Anti-Cardiac Troponin I (ab188877, 1:200), or Anti-PGP9.5 (ab108986, 1:200). Jackson ImmunoResearch: F(ab')2 Anti-Chicken 488 (703-546-155, 1:500) and Anti-Rabbit 647 (711-606-152, 1:500).

Validation	Antibody validation was performed by comparing immunofluorescence signal with established literature. Antibodies have also been previously validated (e.g. vendor, other labs).

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	C57BL/6J mice (Jackson Laboratories, Strain #000664, 4-12 week male or female). Fos2A-iCreER (TRAP2; JAX #030323) and B6;129S6-Gt(ROSA)26Sortm14(CAG-tdTomato)/Hze/J (Ai14; JAX #007908) were crossed to generate double transgenicTRAP2 mice (4-12 week male or female mice were used for this study).
Wild animals	No wild animals were used in this study.
Reporting on sex	Analysis on both sexes were reported in this study.
Field-collected samples	No field samples were used in this study.
Ethics oversight	All experimental protocols were approved by the Stanford University Institutional Animal Care and Use Committee (IACUC) following the National Institutes of Health guidelines for the care and use of laboratory animals.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Article

Coordination of bacterial cell wall and outer membrane biosynthesis

https://doi.org/10.1038/s41586-023-05750-0

Received: 16 July 2021

Accepted: 23 January 2023

Published online: 1 March 2023

Open access

Check for updates

Katherine R. Hummels[1], Samuel P. Berry[2], Zhaoqi Li[1], Atsushi Taguchi[1,3], Joseph K. Min[2], Suzanne Walker[1], Debora S. Marks[2] & Thomas G. Bernhardt[1,4] ✉

Gram-negative bacteria surround their cytoplasmic membrane with a peptidoglycan (PG) cell wall and an outer membrane (OM) with an outer leaflet composed of lipopolysaccharide (LPS)[1]. This complex envelope presents a formidable barrier to drug entry and is a major determinant of the intrinsic antibiotic resistance of these organisms[2]. The biogenesis pathways that build the surface are also targets of many of our most effective antibacterial therapies[3]. Understanding the molecular mechanisms underlying the assembly of the Gram-negative envelope therefore promises to aid the development of new treatments effective against the growing problem of drug-resistant infections. Although the individual pathways for PG and OM synthesis and assembly are well characterized, almost nothing is known about how the biogenesis of these essential surface layers is coordinated. Here we report the discovery of a regulatory interaction between the committed enzymes for the PG and LPS synthesis pathways in the Gram-negative pathogen *Pseudomonas aeruginosa*. We show that the PG synthesis enzyme MurA interacts directly and specifically with the LPS synthesis enzyme LpxC. Moreover, MurA was shown to stimulate LpxC activity in cells and in a purified system. Our results support a model in which the assembly of the PG and OM layers in many proteobacterial species is coordinated by linking the activities of the committed enzymes in their respective synthesis pathways.

The biosynthetic pathways for phospholipids, PG and LPS in Gram-negative bacteria rely on shared precursor pools (Fig. 1a). Therefore, flux through each pathway must be balanced to prevent overconsumption of essential precursors by a single pathway[4–6]. LPS biosynthesis requires both UDP-*N*-acetylglucosamine (UDP-GlcNAc) and acyl-ACP molecules that are also used for PG and phospholipid biosynthesis, respectively[6,7]. Additionally, overproduction of LPS results in the toxic accumulation of LPS intermediates in the inner membrane[8]. Flux through the LPS pathway must therefore be tightly regulated.

Enterobacteria such as *Escherichia coli* control LPS synthesis through regulated proteolysis of the committed enzyme, *Ec*LpxC, by FtsH[4,9]. Previous studies of LpxC in *P. aeruginosa* (*Pa*LpxC), however, showed that it was not proteolysed[10]. Accordingly, N-terminally His-tagged *Pa*LpxC (H–*Pa*LpxC) accumulated normally in an *ftsH*-deletion mutant (Extended Data Fig. 1a). Thus, LPS biogenesis in *P. aeruginosa* seems to be regulated through a mechanism distinct from that in enterobacteria.

*Pa*LpxC is activated by *Pa*MurA

The overproduction of H–*Ec*LpxC but not H–*Pa*LpxC inhibited growth of *P. aeruginosa* and increased cellular levels of LPS (Extended Data Figs. 1b and 2a), suggesting that *Pa*LpxC is regulated in *P. aeruginosa* cells through a mechanism ineffective against *Ec*LpxC. To identify possible regulatory factors, H–*Pa*LpxC or H–*Ec*LpxC was produced in *P. aeruginosa* and interaction partners were identified following affinity purification. *Pa*MurA and PA4701 were the only proteins enriched with H–*Pa*LpxC but not H–*Ec*LpxC (Supplementary Table 1). PA4701 is a non-essential protein (Extended Data Fig. 3a) of unknown function, whereas *Pa*MurA is the essential committed enzyme for PG synthesis[11,12] (Fig. 1a and Extended Data Fig. 3b,c). We focused on the *Pa*MurA–*Pa*LpxC interaction and validated it using in vivo pulldown assays with H–*Pa*LpxC and N-terminally Flag-tagged *Pa*MurA (F–*Pa*MurA; Extended Data Fig. 4a). Notably, the co-purification was enhanced when cells were treated with the LpxC inhibitor CHIR-090, suggesting that the drug-bound LpxC enzyme may have a greater affinity for MurA (Extended Data Fig. 4a). Purified His-tagged *Pa*MurA (H–*Pa*MurA; Extended Data Fig. 5a) was also specifically pulled down by Flag-tagged *Pa*LpxC (F–*Pa*LpxC) using anti-Flag resin, indicating that the interaction was direct (Extended Data Fig. 4b). Purified F–*Pa*LpxC and H–*Pa*MurA were subjected to size-exclusion chromatography (SEC) individually or as 1:1 mixtures with or without CHIR-090. In the mixed sample, a large proportion of protein eluted at a volume corresponding to the heterodimer of F–*Pa*LpxC and H–*Pa*MurA (Fig. 1b). Consistent with the pulldown assays, re-application of the heterodimer fractions to the SEC column showed that the complex remained most stable in the presence of CHIR-090 (Fig. 1b). Notably, H–*Pa*MurA stimulated the enzymatic activity of F–*Pa*LpxC (Fig. 1c). This stimulation was specific to *Pa*MurA as the addition of H–*Ec*MurA had no effect (Fig. 1c). We conclude that *Pa*MurA is a direct and specific activator of *Pa*LpxC in vitro.

[1]Department of Microbiology, Blavatnik Institute, Harvard Medical School, Boston, MA, USA. [2]Department of Systems Biology, Blavatnik Institute, Harvard Medical School, Boston, MA, USA. [3]SANKEN (The Institute of Scientific and Industrial Research), Osaka University, Ibaraki, Japan. [4]Howard Hughes Medical Institute, Boston, MA, USA. ✉e-mail: thomas_bernhardt@hms.harvard.edu

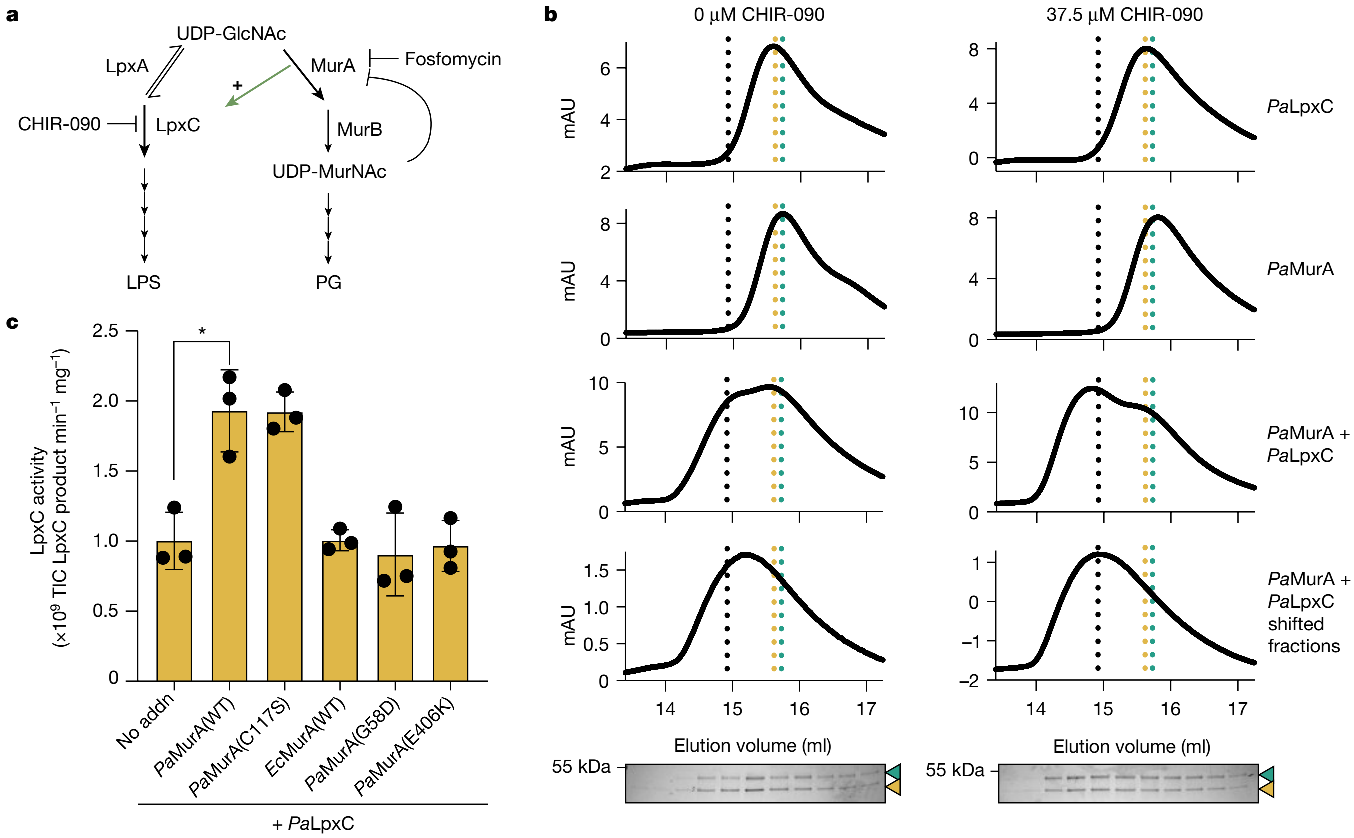


Fig. 1 | *Pa*MurA interacts with and activates *Pa*LpxC. a, Schematic representation of the biosynthetic pathways responsible for PG and LPS biosynthesis, showing the committed enzymes MurA and LpxC, and other relevant enzymes (LpxA and MurB). T-bars indicate inhibition by the antibiotics fosfomycin and CHIR-090. The green arrow indicates activation of LpxC by MurA. **b**, Size-exclusion chromatography in which 7.5 µM of purified F–LpxC and H–MurA were resolved either individually or as a mixture in the presence or absence of 37.5 µM CHIR-090 as indicated. The shifted F–LpxC + H–MurA fractions were subsequently collected, resubjected to size-exclusion chromatography, and the resulting fractions were resolved by SDS–PAGE followed by Coomassie staining. Dotted lines indicate the peak mobilities for F–LpxC (gold), H–MurA (cyan) or the shifted F–LpxC + H–MurA fractions (black) in the presence of CHIR-090. The mobilities of H–MurA and F–LpxC in SDS–PAGE assays are indicated by a cyan or gold arrowhead, respectively. Data are representative of two replicates. For gel source data, see Supplementary Fig. 1. **c**, Catalytic activity (total ion counts (TIC)) of purified *Pa*LpxC (100 nM) alone (no addition (No addn)) or in the presence of MurA variants (100 nM). Dots indicate the values obtained for three individual replicates, bars indicate the mean, and error bars represent their standard deviation. *P = 0.0109 (unpaired, two-tailed t-test).

We reasoned that if *Pa*MurA is an essential activator of *Pa*LpxC in *P. aeruginosa* cells, then *Pa*MurA depletion should result in a phenotype that resembles the simultaneous inactivation of both essential enzymes. Cells treated with the MurA inhibitor fosfomycin have a PG synthesis inhibition phenotype involving membrane bleb formation and lysis[11] (Extended Data Fig. 3d–f). However, cells depleted of *Pa*MurA instead adopted an enlarged, ovoid shape (Extended Data Fig. 3e,f). This phenotype resembled that of cells treated with both CHIR-090 and fosfomycin (Extended Data Fig. 3d–f), suggesting that *Pa*MurA depletion impairs the synthesis of both PG and LPS. Accordingly, *Pa*MurA depletion reduced LPS levels, whereas depletion of the next enzyme in the pathway, *Pa*MurB, did not (Extended Data Figs. 2b and 3b,c) and instead caused a terminal phenotype resembling that following fosfomycin treatment (Extended Data Fig. 3d–f). We conclude that *Pa*MurA is required for the biosynthesis of normal cellular levels of LPS.

*Pa*MurA has two essential functions

Overproduction of wild-type (WT) *Pa*MurA did not increase LPS levels as expected for an activator of *Pa*LpxC (Extended Data Fig. 2c). We suspected this result might be due to the enzymatic activity of *Pa*MurA competing with the LPS pathway for UDP-GlcNAc (Fig. 1a) thereby preventing runaway LPS synthesis. Alternatively, a particular MurA conformer[13] may be the activator and it may not be produced in sufficient levels to stimulate LPS synthesis on *Pa*MurA overexpression. These scenarios predict that the overproduction of catalytically inactive or conformationally trapped *Pa*MurA variants should cause lethal levels of LPS production. We therefore searched for toxic *Pa*MurA variants. *P. aeruginosa* was transformed with a plasmid encoding mutagenized Pa*murA* under the control of an isopropyl-β-D-thiogalactoside (IPTG)-regulated promoter. The resulting library was pooled and grown in liquid medium with inducer. Plasmids that caused lysis were purified from the culture supernatant, and the Pa*murA* genes from those causing an IPTG-dependent growth defect on retransformation were sequenced.

Twenty-three toxic *Pa*MurA variants (designated *Pa*MurA*) were identified (Fig. 2a and Extended Data Fig. 6a). Their growth inhibitory activity was alleviated by a normally lethal concentration of CHIR-090, suggesting that they hyperactivate *Pa*LpxC (Extended Data Fig. 6a). The amino acid changes in the *Pa*MurA* variants mapped around the active site of the enzyme[14] (Extended Data Fig. 6b) and included the catalytic cysteine residue (C117) that forms a covalent intermediate with the phosphoenolpyruvate substrate[15]. A subset of *Pa*MurA* variants were purified (Extended Data Fig. 5a) and found to have markedly reduced enzymatic activity while retaining their ability to activate purified *Pa*LpxC (Extended Data Fig. 6c,d).

The equivalent of the C117S substitution in *Pa*MurA has been well characterized in other orthologues in which it has been shown to trap the enzyme in a closed conformation bound to its product[16,17].

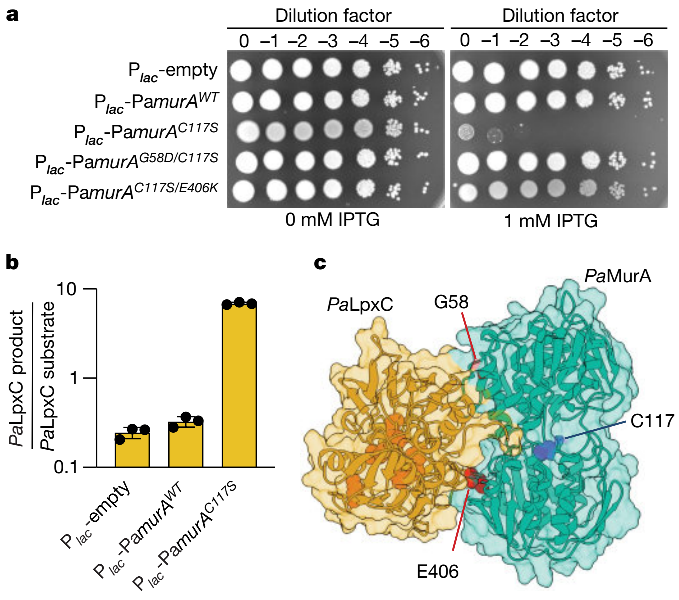


Fig. 2 | *Pa*MurA(C117S) is a potent activator of *Pa*LpxC in vivo. a, Viability assay in which serial dilutions of PAO1 harbouring an empty plasmid or the indicated Pa*murA* variant under IPTG-inducible control were plated on LB agar with or without IPTG supplementation as indicated. Data are representative of three biological replicates. **b**, Ratio of LpxC product to substrate detected by liquid chromatography–mass spectrometry in the aqueous fraction of methanol–chloroform-extracted *P. aeruginosa* whole-cell lysates. Dots indicate the values obtained for three biological replicates, bars indicate the mean, and error bars represent their standard deviation. **c**, Model structure of the *Pa*LpxC–*Pa*MurA complex predicted by AlphaFold[19]. *Pa*LpxC is represented in gold and *Pa*MurA is represented in cyan. *Pa*MurA residues G58 and E406 are highlighted in red and residue C117 is highlighted in navy. *Pa*LpxC active-site residues[34] are depicted in orange.

We therefore chose to study the potential in vivo activation of *Pa*LpxC by *Pa*MurA(C117S) further. The LpxC substrate (UDP-3-*O*-(*R*-3-hydroxydecanoyl)-*N*-acetylglucosamine) and product (UDP-3-*O*-(*R*-3-hydroxydecanoyl)-glucosamine) were extracted and quantified from cells overproducing *Pa*MurA(WT) or *Pa*MurA(C117S). The *Pa*LpxC product/substrate ratio was unchanged in cells overproducing *Pa*MurA(WT) relative to an empty vector control (Fig. 2b and Extended Data Fig. 7). However, overproduction of *Pa*MurA(C117S) led to a 100-fold increase in the *Pa*LpxC product/substrate ratio and increased LPS levels, indicating that this variant is a potent activator of *Pa*LpxC in vivo (Fig. 2b and Extended Data Figs. 2c and 7). We therefore conclude that *Pa*MurA activates *Pa*LpxC in cells but that only catalytically inactive *Pa*MurA variants promote toxic levels of LPS synthesis when they are overproduced owing to their inability to compete with the LPS synthesis pathway for UDP-GlcNAc.

To further characterize the interaction between *Pa*LpxC and *Pa*MurA, we searched for non-toxic derivatives of *Pa*MurA(C117S), reasoning that some would be unable to bind *Pa*LpxC. Survivors of F–*Pa*MurA(C117S) production from a mutagenized plasmid were selected followed by an immunoblot-based screen for isolates producing stable, full-length F–*Pa*MurA(C117S). This procedure identified *Pa*MurA(C117S/G58D) and *Pa*MurA(C117S/E406K). We confirmed that both variants were not toxic when overproduced to levels similar to *Pa*MurA(C117S) (Fig. 2a and Extended Data Fig. 8a). We next purified H–*Pa*MurA(G58D) and H–*Pa*MurA(E406K) variants with intact active sites (Extended Data Fig. 5a). Both retained MurA activity (Extended Data Fig. 8b), and consistent with the genetic results, they failed to activate F–*Pa*LpxC in vitro (Fig. 1c). Notably, however, whereas H–*Pa*MurA(G58D) was unable to bind to F–*Pa*LpxC, H–*Pa*MurA(E406K) exhibited binding activity similar to that of H–*Pa*MurA(WT) (Extended Data Fig. 8c). Thus, the G58D change impairs the ability of *Pa*MurA to bind *Pa*LpxC, whereas the E406K substitution seems to disrupt the activation mechanism following complex formation. The residues G58 and E406 both lie on the same surface of the *Pa*MurA structure, suggesting that this face comprises the *Pa*LpxC-binding interface (Fig. 2c).

The results thus far suggest that *Pa*MurA may have two essential functions: PG synthesis and activation of *Pa*LpxC. This model predicts that a catalytically active variant of *Pa*MurA that cannot activate *Pa*LpxC should fail to complement a Pa*murA* deletion but remain capable of complementing a catalytically dead *Pa*MurA variant that retains its *Pa*LpxC activation function. To test this prediction, Pa*murA*^{WT} was placed under P_{lac} control and the native allele was either deleted or converted to Pa*murA*^{C117S}. We then introduced a plasmid with or without the Pa*murA*^{G58D} gene controlled by an arabinose-inducible promoter (P_{ara}). As expected, *Pa*MurA(WT) depletion resulted in a severe growth defect in either the Δ*murA* or Pa*murA*^{C117S} backgrounds (Extended Data Fig. 8d). Notably, however, the production of *Pa*MurA(G58D) was able to restore growth on *Pa*MurA(WT) depletion in the context of the Pa*murA*^{C117S} allele, but not the *murA* deletion (Extended Data Fig. 8d). We therefore conclude that the *Pa*LpxC activation function of *Pa*MurA is essential and separable from its catalytic activity.

LpxC–MurA interaction is conserved

To investigate the conservation of the LpxC–MurA regulatory interaction, we used evolutionary covariation analysis[18]. However, because both enzymes are conserved throughout Gram-negative bacteria but only a subset are likely to interact, we could not reliably detect residues that covary between LpxC and MurA without first knowing which regions of the two proteins interact; non-interacting pairs among the genomes analysed generated too much 'noise' to detect a clear interaction signature. We therefore modelled the structure of the LpxC–MurA complex with AlphaFold[19], which predicted a high-confidence structure for *Pa*LpxC–*Pa*MurA but not between the corresponding *E. coli* proteins (Fig. 2c and Extended Data Fig. 9a,b). The G58 and E406 residues of *Pa*MurA implicated in *Pa*LpxC binding and/or activation lie at the modelled interaction interface (Fig. 2c), supporting the accuracy of the structure prediction. Using this model as a guide, evolutionary covariation analysis of residues predicted to be at the interaction interface was carried out on LpxC–MurA pairs from 8,302 proteobacterial genomes. Each LpxC–MurA pair was then assigned an 'interaction score' as an indication of how strongly the two proteins were predicted to interact (Fig. 3a).

The *P. aeruginosa* LpxC–MurA pair had a high interaction score as expected, as did those from other Pseudomonadales and a subset of other gammaproteobacteria including the Legionalles, Xanthomonadales and Oceanospirillales orders (Fig. 3a). In addition, high interaction scores were also observed in a subset of Rhodospirillales, an order of alphaproteobacteria that is highly divergent from *P. aeruginosa* (Fig. 3a). On the basis of the distribution of interaction scores, we estimate that 48% of gammaproteobacteria and 35% of alphaproteobacteria encode LpxC–MurA pairs that are capable of interacting (Extended Data Fig. 9e). To investigate the accuracy of the covariation analysis, three additional LpxC–MurA pairs with high interaction scores (*Magnetospirillum kuznetsovii*, *Xanthomonadales* spp. and *Legionella pneumophila*) and two LpxC–MurA pairs with low interaction scores (*E. coli* and *Acinetobacter baumannii*) were purified (Extended Data Fig. 5a) and their ability to interact was tested in vitro. Only those pairs with high interaction scores were observed to interact (Fig. 3b). Furthermore, *L. pneumophila* LpxC (*Lpn*LpxC), but not *Ec*LpxC, exhibited increased activity in vitro in the presence of its cognate MurA (Extended Data Fig. 9f,g). Thus, we conclude that LpxC and MurA from a variety of proteobacteria interact, suggesting that the regulatory link between the PG and LPS biogenesis pathways is conserved in a variety of Gram-negative bacteria.

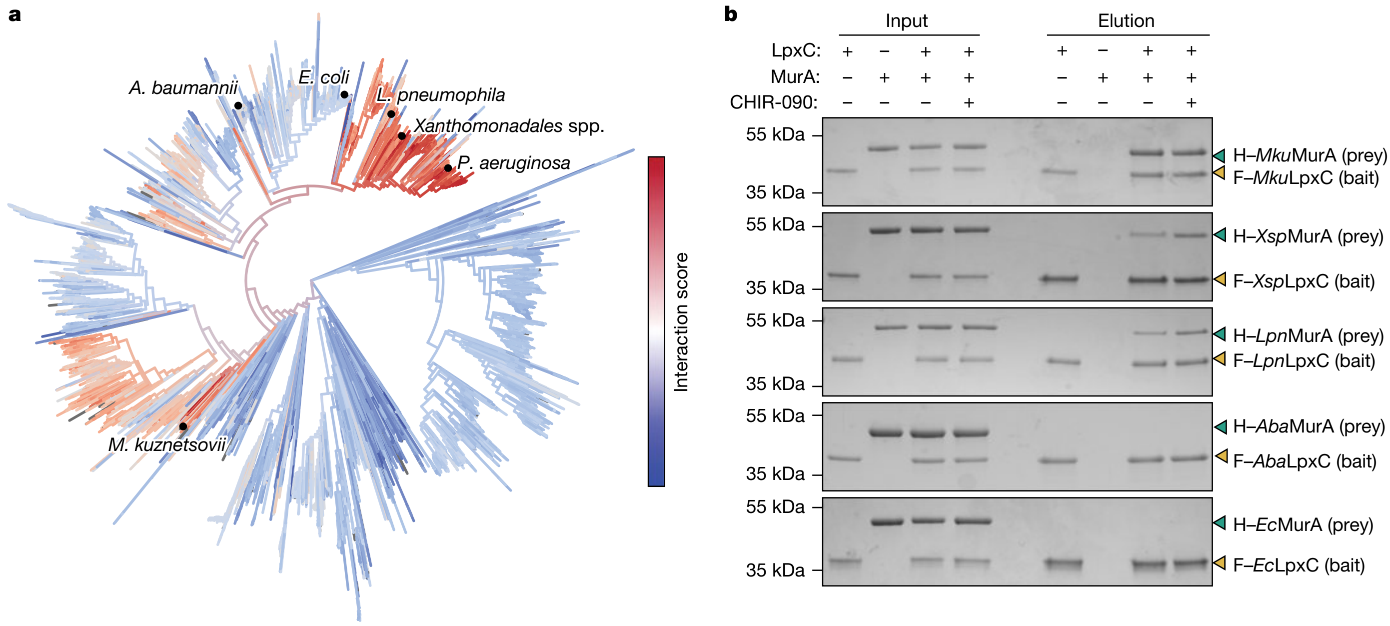


Fig. 3 | Direct interaction between LpxC and MurA is observed in diverse bacteria. a, MurA–LpxC interaction scores calculated from a direct coupling analysis-based approach (Methods) plotted on a maximum-likelihood phylogenetic tree based on concatenated MurA and LpxC sequences generated with FastTree. Experimentally tested interactions are highlighted. **b**, Purified F–LpxC (2.5 μM) was mixed in a 1:1 ratio with purified H–MurA in the presence or absence of CHIR-090 (5.7 μM) as indicated. The mixtures were pulled down with anti-Flag resin and the input and eluate were subjected to SDS–PAGE and Coomassie staining. Mobilities of H–MurA and F–LpxC are indicated by a cyan or gold arrowhead, respectively. Data are representative of at least two replicates. For gel source data, see Supplementary Fig. 1.

Model for balanced LPS and PG synthesis

The LPS biosynthetic pathway shares critical precursors with both the phospholipid and PG biosynthetic pathways (Fig. 1a). In *E. coli* and other enterobacteria, the need for balanced consumption of acyl-ACP molecules by the LPS and phospholipid biosynthetic pathways has been well documented[4,20–23]. By contrast, little is known about how UDP-GlcNAc utilization is coordinated by the LPS and PG biosynthetic pathways to allow uniform cell envelope expansion. We propose that the *Pa*LpxC-activating function of *Pa*MurA serves to limit LPS biosynthesis such that it cannot outrun PG biosynthesis. In support of this model, overexpression of *Pa*MurA(WT), which is capable of competing with the LPS synthesis pathway for UDP-GlcNAc, had no impact on cellular viability (Fig. 2a). Overexpression of the catalytically inactive variant *Pa*MurA(C117S), however, resulted in the imbalanced activation of *Pa*LpxC in vivo, resulting in cell death (Fig. 2a). By limiting the catalytically active population of *Pa*LpxC to *Pa*MurA levels, runaway LPS biosynthesis is not possible. On the other hand, MurA has been shown to be feedback inhibited by the downstream PG precursor UDP-MurNAc[24]. Under conditions in which the flux of PG precursor synthesis is too high, UDP-MurNAc will accumulate such that it binds to MurA and locks it into a closed conformation similar to that of *Pa*MurA(C117S)[13,16]. Thus, when PG biosynthesis outpaces LPS biosynthesis, MurA will be subject to feedback inhibition, decreasing its catalytic activity without affecting its ability to activate *Pa*LpxC to stimulate LPS biosynthesis and rebalance the pathways.

In enterobacteria, LpxC is proteolysed by FtsH, which is in turn regulated by at least two accessory proteins: LapB and YejM[5,8,25,26]. Although FtsH is broadly conserved, LapB and YejM are observed only in a subset of proteobacterial genomes. By contrast, LpxC and MurA are nearly universally conserved in proteobacteria. Our computational analysis suggests that a substantial group of gammaproteobacteria and alphaproteobacteria control LpxC through activation by MurA (Fig. 3a). Why bacteria use different strategies to regulate flux through the LPS biosynthetic pathway is unclear. We suggest, however, that it may reflect differences in their environmental niche. For example, *P. aeruginosa* preferentially utilizes tricarboxylic acids over carbohydrates as a carbon source[27]. Thus, activated sugars such as UDP-GlcNAc may be more limiting for *P. aeruginosa* than for sugar-loving *E. coli* such that it focuses its regulation of LpxC on the appropriate partitioning of UDP-GlcNAc. It would not be surprising if future studies uncover other as-yet-unknown strategies of LpxC regulation in other organisms.

Conclusions

In conclusion, our work has identified a previously unknown and unexpected regulatory interaction between essential enzymes involved in the production of two different layers of the Gram-negative cell envelope. In addition to uncovering a potential mechanism for coordinating the biogenesis of the cell wall and OM, these findings also have implications for antibiotic development. MurA is the target of the antibiotic fosfomycin and LpxC has been the subject of several drug discovery campaigns that have yielded inhibitors with potent antibacterial activity[11,28–33]. The connection between these enzymes identified here suggests that it may be possible to develop dual-targeting drugs that alter both MurA and LpxC activity to simultaneously disrupt PG and OM assembly to kill *P. aeruginosa* and/or sensitize it to other antibiotics made ineffective by the barrier function of its envelope. These findings also raise the possibility that there are many more regulatory interactions between enzymes involved in the biogenesis of different cell envelope components waiting to be discovered.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-05750-0.

1. Silhavy, T. J., Kahne, D. & Walker, S. The bacterial cell envelope. *Cold Spring Harb. Perspect. Biol.* **2**, a000414 (2010).
2. Zgurskaya, H. I., Lopez, C. A. & Gnanakaran, S. Permeability barrier of Gram-negative cell envelopes and approaches to bypass it. *ACS Infect. Dis.* **1**, 512–522 (2015).
3. Bush, K. Antimicrobial agents targeting bacterial cell walls and cell membranes. *Rev. Sci. Tech.* **31**, 43056 (2012).
4. Ogura, T. et al. Balanced biosynthesis of major membrane components through regulated degradation of the committed enzyme of lipid A biosynthesis by the AAA protease FtsH (HflB) in *Escherichia coli*. *Molec. Microbiol.* **31**, 833–844 (1999).
5. Klein, G., Kobylak, N., Lindner, B., Stupak, A. & Raina, S. Assembly of lipopolysaccharide in *Escherichia coli* requires the essential LapB heat shock protein. *J. Biol. Chem.* **289**, 14829–14853 (2014).
6. Raetz, C. R. Biochemistry of endotoxins. *Annu. Rev. Biochem.* **59**, 129–170 (1990).
7. Anderson, M. S., Bulawa, C. E. & Raetz, C. R. The biosynthesis of gram-negative endotoxin. Formation of lipid A precursors from UDP-GlcNAc in extracts of *Escherichia coli*. *J. Biol. Chem.* **260**, 15536–15541 (1985).
8. Guest, R. L., Same Guerra, D., Wissler, M., Grimm, J. & Silhavy, T. J. YejM modulates activity of the YciM/FtsH protease complex to prevent lethal accumulation of lipopolysaccharide. *mBio* **11**, e00598-20 (2020).
9. Anderson, M. S. et al. UDP-*N*-acetylglucosamine acetyltransferase of *Escherichia coli*: the first step in endotoxin biosynthesis is thermodynamically unfavorable. *J. Biol. Chem.* **268**, 19858–19865 (1993).
10. Langklotz, S., Schakermann, M. & Narberhaus, F. Control of lipopolysaccharide biosynthesis by FtsH-mediated proteolysis of LpxC is conserved in enterobacteria but not in all gram-negative bacteria. *J. Bacteriol.* **193**, 1090–1097 (2011).
11. Kahan, F. M., Kahan, J. S., Cassidy, P. J. & Kropp, H. The mechanism of action of fosfomycin (phosphonomycin). *Ann. N. Y. Acad. Sci.* **235**, 364–386 (1974).
12. Strominger, J. L. Enzymic transfer of pyruvate to uridine diphosphoacetylglucosamine. *Biochem. Biophys. Acta* **30**, 645–646 (1958).
13. Zhu, J. Y. et al. Functional consequence of covalent reaction of phosphoenolpyruvate with UDP-N-acetylglucosamine 1-carboxyvinyltransferase (MurA). *J. Biol. Chem.* **287**, 12657–12667 (2012).
14. Eschenberg, S. & Schonbrunn, E. Comparative X-ray analysis of the un-liganded fosfomycin-target MurA. *Proteins* **40**, 290–298 (2000).
15. Wanke, C. & Amrhein, N. Evidence that the reaction of the uDP-N-acetylglucosamine 1-carboxyvinyltransferase proceeds through the O-phosphothioketal of pyruvic acid bound to Cys115 of the enzyme. *Eur. J. Biochem.* **218**, 861–870 (1993).
16. Eschenburg, S., Priestman, M. & Schönbrunn, E. Evidence that the fosfomycin target Cys115 in UDP-N-acetylglucosamine enolpyruvyl transferase (MurA) is essential for product release. *J. Biol. Chem.* **280**, 3757–3763 (2005).
17. Schönbrunn, E., Eschenberg, S., Krekel, F., Luger, K. & Amrhein, N. Role of the loop containing residue 115 in the induced-fit mechanism of the bacterial cell wall enzyme MurA. *Biochemistry* **39**, 2164–2173 (2000).
18. Hopf, T. A. et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **3**, e03430 (2014).
19. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
20. Mohan, S., Kelly, T. M., Eveland, S. S., Raetz, C. R. & Anderson, M. S. An *Escherichia coli* gene (FabZ) encoding (3R)-hydroxymyristoyl acyl carrier protein dehydrase. Relation to fabA and suppression of mutations in lipid A biosynthesis. *J. Biol. Chem.* **269**, 32896–32903 (1994).
21. Thomanek, N. et al. Intricate crosstalk between lipopolysaccharide, phospholipid and fatty acid metabolism in *Escherichia coli* modulates proteolysis of LpxC. *Front. Microbiol.* **9**, 3285 (2018).
22. Mostafavi, M. et al. Interplay of *Klebsiella pneumoniae* fabZ and lpxC mutations leads to LpxC inhibitor-dependent growth resulting from loss of membrane homeostasis. *mSphere* **3**, e00508-18 (2018).
23. Zeng, D. et al. Mutants resistant to LpxC inhibitors by rebalancing cellular homeostasis. *J. Biol. Chem.* **288**, 5475–5486 (2013).
24. Mizyed, S. et al. UDP-*N*-acetylmuramic acid (UDP-MurNAc) is a potent inhibitor of MurA (enolpyruvyl-UDP-GlcNAc synthase). *Biochemistry* **44**, 4011–4017 (2005).
25. Fivenson, E. M. & Bernhardt, T. G. An essential membrane protein modulates the proteolysis of LpxC to control lipopolysaccharide synthesis in *Escherichia coli*. *mBio* **11**, e00939-20 (2020).
26. Clairfeuille, T. et al. Structure of the essential inner membrane lipopolysaccharide–PbgA complex. *Nature* **584**, 479–483 (2020).
27. Collier, D. N., Hager, P. W. & Phibbs, P. V. Jr Catabolite repression control in the *Pseudomonads*. *Res. Microbiol.* **147**, 551–561 (1996).
28. Kline, T. Potent, novel in vitro inhibitors of the *Pseudomonas aeruginosa* deacetylase LpxC. *J. Med. Chem.* **45**, 3112–3129 (2002).
29. McClerren, A. L. et al. A slow, tight-binding inhibitor of the zinc-dependent deacetylase LpxC of Lipid A biosynthesis with antibiotic activity comparable to ciprofloxacin. *Biochemistry* **44**, 16574–16583 (2005).
30. Langsdorf, E. F. et al. Screening for antibacterial inhibitors of the UDP-3-*O*-(*R*-3-hyd roxymyristoyl)-*N*-acetylglucosamine deacetylase (LpxC) using a high-throughput mass spectrometry assay. *J. Biomol. Screen.* **15**, 52–61 (2010).
31. Erwin, A. L. Antibacterial drug discovery targeting the lipopolysaccharide biosynthetic enzyme LpxC. *Cold Spring Harb. Perspect. Med.* **6**, a025304 (2016).
32. Krause, K. M. et al. Potent LpxC inhibitors with *in vitro* activity against multidrug-resistant *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* **63**, e00977-19 (2019).
33. Ushiyama, F. et al. Lead optimization of 2-hydroxymethyl imidazoles as non-hydroxamate LpxC inhibitors: discovery of TP0586532. *Bioorg. Med. Chem.* **30**, 115964 (2021).
34. Hernick, M. & Fierke, C. A. Catalytic mechanism and molecular recognition of *E. coli* UDP-3-*O*-(*R*-3-hydroxymyristoyl)-*N*-acetylglucosamine deacetylase probed by mutagenesis. *Biochemistry* **45**, 15240–15248 (2006).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

Methods

Media, bacterial strains and plasmids

As indicated, cells were grown in LB (1% tryptone, 0.5% yeast extract, 0.5% NaCl) or minimal M9 medium[35] supplemented with 0.2% casamino acids and 0.2% glucose. The following concentrations of antibiotics were used: chloramphenicol, 25 μg ml^{-1}; kanamycin, 25 μg ml^{-1}; gentamycin, 30 μg ml^{-1} (*P. aeruginosa*) or 15 μg ml^{-1} (*E. coli*); carbenicillin, 200 μg ml^{-1} (*P. aeruginosa*) or 50 μg ml^{-1} (*E. coli*). Plasmids used in this study are listed in Supplementary Table 2. The bacterial strains used in this study are listed in Supplementary Table 3. All *P. aeruginosa* strains used in the reported experiments are derivatives of PAO1. Primers used in this study are listed in Supplementary Table 4. Unless stated otherwise, polymerase chain reaction (PCR) was carried out using Q5 polymerase (NEB M0492L) for cloning purposes and GoTaq Green DNA polymerase (Promega M7123) for diagnostic purposes, both according to the manufacturer's instructions. Plasmid DNA and PCR fragments were purified using the PurePlasmid miniprep kit (CW Biosciences CW0500M) or the DNA Clean-up kit (CW Biosciences CW2301M), respectively. Details on strain and plasmid construction can be found in the Supplementary Information. No statistical methods were used to predetermine sample sizes used for experiments, but sample sizes are in line with field standards.

Isolation of toxic *murA* alleles

Alleles of *murA* that are toxic when overexpressed were isolated as described previously[36] with slight modifications. First, the *murA* gene was mutagenized by PCR with Taq polymerase (NEB M0267L) from purified pKH37 plasmid using primer pair 15/76 in five separate reactions. A 500 ng quantity of each of the five resulting *murA** pools was then separately used as a megaprimer to amplify 50 ng of pKH37 using Q5 polymerase (NEB M0492L) in 50-μl reactions with the following thermocycler settings: 1) 95 °C for 3 min; 2) 95 °C for 50 s; 3) 60 °C for 50 s; 4) 72 °C for 10 min; 5) repeat steps 2–4 for a total of 25 cycles. Twenty units of DpnI (NEB R0176L) was subsequently added to each reaction and digestion of unamplified DNA was allowed to proceed at 37 °C for 1.5 h. Each reaction was drop dialysed using 0.025-μm mixed cellulose ester membranes (Millipore VSWP02500) floated on Milli-Q water for 20 min. A 12 μl volume of each dialysed product was then separately transformed into 100 μl of NEB-5-alpha electrocompetent *E. coli* (NEB C2989K), and recovered in SOC outgrowth medium (NEB B9020S), and transformants were selected by plating the outgrowth onto LB agar supplemented with gentamycin. Approximately 1 million colonies from each of the five libraries were separately pooled and the pKH37 derivatives were purified using the PurePlasmid miniprep kit (CW Biosciences CW0500M).

PAO1 was grown in 25 ml LB overnight at 37 °C, centrifuged at 12,000*g* for 5 min, and resuspended in an equal volume of 300 mM sucrose. The centrifugation and resuspension steps were repeated for a total of four washes. After a final centrifugation at 12,000*g* for 5 min, the cell pellet was resuspended in 1/20 of the original volume. A 1.2 μg quantity of pKH37-derived libraries was separately transformed into 150 μl of electrocompetent PAO1. Transformation reactions were recovered in LB at 37 °C for 1 h and transformants were selected by plating the outgrowth onto LB agar supplemented with 30 μg ml^{-1} gentamycin. Approximately 1–2 million colonies from each library were separately pooled.

To identify pKH37 variants that harbour a toxic allele of *murA*, each PAO1 × pKH37* library was diluted to an optical density at 600 nm (OD_{600nm}) of 0.01 in 3 ml LB supplemented with 30 μg ml^{-1} gentamycin and grown at 37 °C for 2 h 45 min. IPTG was added to each culture to a final concentration of 1 mM and the incubation at 37 °C was continued for an additional 2 h. Each culture was then subjected to centrifugation at 21,130*g* for 5 min and released DNA from each library was separately purified from 2.5 ml of the supernatant using the DNA clean-up kit (CW Biosciences CW2301M). The purified DNA was subsequently re-transformed into PAO1 electrocompetent cells prepared as described above and transformants were selected by plating the outgrowth on LB agar supplemented with gentamycin. The resulting colonies were then patched onto LB agar plates supplemented with gentamycin with and without 1 mM IPTG and grown overnight at 37 °C. The *murA* allele encoded by IPTG-sensitive isolates was determined by Sanger sequencing using primer pair 15/76.

Isolation of MurA(C117S) loss-of-toxicity variants

pKH100 was mutagenized by passaging it through *E. coli* strain XL1-red. pKH100 was transformed into XL1-red and five separate cultures were inoculated, each using four unique transformants. After propagation overnight, plasmids were purified using the PurePlasmid miniprep kit (CW Biosciences CW0500M) to create five pKH100* libraries. PAO1 was grown in 25 ml LB overnight at 37 °C, centrifuged at 12,000*g* for 5 min, and resuspended in an equal volume of 300 mM sucrose. The centrifugation and resuspension steps were repeated for a total of four washes. After a final centrifugation at 12,000*g* for 5 min, the cell pellet was resuspended in 1/20 of the original volume. A 1.2 μg quantity of each pKH100* was separately transformed into 150 μl of electrocompetent PAO1. Transformation reactions were recovered in LB at 37 °C for 1 h and transformants were selected by plating the outgrowth onto LB agar supplemented with gentamycin. Approximately 1–10 million colonies from each library were separately pooled and stored in LB + 10% dimethylsulfoxide (DMSO) at −80 °C.

PAO1 × pKH100* libraries were plated on LB agar supplemented with 30 μg ml^{-1} gentamycin and 1 mM IPTG, which allowed growth of only approximately 0.3–1% of all colony-forming units. A total of 17 or 18 IPTG-resistant colonies from each pool were streak-purified and the expression level and molecular weight of F–MurA(C117S)* from each isolate were determined by western blot analysis after growth in LB and induction for 1 h in mid-log phase. Only those isolates with expression levels and molecular weights consistent with F–MurA(C117S) were subjected to Sanger sequencing with primers 15 and 76. In three of the five libraries analysed, *F–murA^{WT}* was recovered. *F–murA$^{C117S/G58D}$* and *F–murA$^{C117S/E406K}$* were each recovered in one of the five libraries.

Protein purification

Details on protein purification can be found in the Supplementary Information.

Viability assays

Overnight cultures grown at 30 °C or 37 °C were centrifuged at 12,000*g* for 2 min and resuspended to an OD_{600nm} of 1 in LB. Resuspensions were serially diluted in LB to 10^{-6} and 5 μl of each dilution was spotted onto LB agar supplemented with the appropriate inducers and antibiotics. Plates were incubated at 37 °C overnight and imaged using a Nikon D3400 camera equipped with a Nikon AF-S micro NIKKOR 40-mm lens.

Western blot analysis

Overnight cultures were diluted to an OD_{600nm} of 0.01 in 3 ml LB, grown at 37 °C for 4 h, and 2 ml was subjected to centrifugation at 12,000*g* for 3 min. When necessary, IPTG was added to 1 mM or arabinose was added to 0.1% 1 h before collecting the cultures. The cell pellet was resuspended to an OD_{600nm} of 20 in 1× SDS sample buffer (50 mM Tris pH 6.8, 10% glycerol, 2% SDS, 0.2% bromophenol blue, 1% β-mercaptoethanol) and boiled at 100 °C for 10 min. The lysate was subsequently sonicated using a Qsonica Q800R3 sonicator with 30 cycles of 1 s on, 1 s off at 25% amplitude. Protein concentration was quantified using the Non-Interfering protein assay kit (G-biosciences 786-005). A 10 μg quantity of protein was loaded onto a 10% polyacrylamide–SDS gel and subjected to electrophoresis at 150 V for 70 min. Protein was transferred to PVDF membranes using the mixed molecular weight protocol on a Bio-Rad Trans-Blot Turbo transfer system and membranes were blocked in TBST (20 mM Tris pH 8.0, 200 mM NaCl, 0.05%

Tween-20) with 5% milk at room temperature for 1 h. Membranes were then probed with anti-His (GenScript A00186-100), anti-Flag (Sigma F7425) or anti-RpoA (BioLegend 663104) diluted 1:3,000, 1:3,000 or 1:100,000, respectively, in TBST + 5% milk at room temperature for 1 h. Membranes were washed three times for 5 min each in TBST and probed with peroxidase-conjugated goat-anti-mouse (Rockland 610-1302) or rabbit TrueBlot: anti-rabbit IgG HRP (Rockland 18-8816-33) diluted 1:3,000 in TBST + 5% milk at room temperature for 1 h. Membranes were washed three times for 5 min each in TBST, developed with SuperSignal West Pico PLUS Chemiluminescent Substrate (Thermo 34580), and imaged using an Azure Biosystems C600 imager.

In vivo co-affinity purification

To identify in vivo interacting partners of *Pa*LpxC in an unbiased manner, two co-affinity purification schemes were used in replicate samples. Specifically, in condition 1, 150 mM NaCl was used in lysis and wash buffers, but in condition 2, 300 mM NaCl was used in lysis and wash buffers. Overnight *P. aeruginosa* cultures were diluted to OD_{600nm} = 0.01 in 500 ml LB and grown at 37 °C to early logarithmic phase. In the case of PA1009, the culture was induced with 0.5% arabinose 1 h before collection. The entire culture was subjected to centrifugation at 13,000*g* for 10 min and the pellets were resuspended in 4 ml of lysis buffer (50 mM Tris pH 7.5, 2% glycerol, 5 mM imidazole, 1% Triton X-100, and 150 mM or 300 mM NaCl) and cells were lysed by sonication with a Q125 Qsonica sonicator. Cell debris was pelleted by centrifugation at 21,130*g* for 15 min at 4 °C and the clarified supernatant was mixed with Ni-NTA resin (Qiagen 30230) rotating end-over-end at 4 °C for 1.5 h. Resin was washed four times with 6 ml of wash buffer (50 mM Tris pH 7.5, 2% glycerol, 25 mM imidazole, 1% Triton X-100, and 150 mM or 300 mM NaCl). Proteins were eluted from the resin in elution buffer (50 mM Tris pH 7.5, 2% glycerol, 250 mM imidazole, 0.1% Triton X-100 and 150 mM NaCl). Eluates were mixed 1:1 with 2× SDS sample buffer (100 mM Tris pH 6.8, 20% glycerol, 4% SDS, 0.4% bromophenol blue), and 40 µl was loaded onto 4–20% polyacrylamide Mini-PROTEAN TGX gels (Bio-Rad 4568094) and subjected to electrophoresis at 80 V for 25 min. The gel was subsequently stained with Coomassie R-250 and entire lanes were excised for mass spectrometry analysis.

Excised lanes were cut into approximately 1-mm^3 pieces. Gel pieces were then subjected to a modified in-gel trypsin digestion procedure[37]. Gel pieces were washed and dehydrated with acetonitrile for 10 min, followed by removal of acetonitrile. Pieces were then completely dried in a speed-vac. Gel pieces were rehydrated with 50 mM ammonium bicarbonate solution containing 12.5 ng μl^{-1} modified sequencing-grade trypsin (Promega) at 4 °C. After 45 min, the excess trypsin solution was removed and replaced with 50 mM ammonium bicarbonate solution to just cover the gel pieces. Samples were then placed in a 37 °C room overnight. Peptides were later extracted by removing the ammonium bicarbonate solution, followed by one wash with a solution containing 50% acetonitrile and 1% formic acid. The extracts were then dried in a speed-vac (about 1 h). The samples were then stored at 4 °C until analysis.

On the day of analysis, the samples were reconstituted in 5–10 µl of high-performance liquid chromatography (HPLC) solvent A (2.5% acetonitrile, 0.1% formic acid). A nanoscale reversed-phase HPLC capillary column was created by packing 2.6-µm C18 spherical silica beads into a fused silica capillary (100 µm inner diameter × about 30 cm length) with a flame-drawn tip[38]. After equilibrating the column, each sample was loaded using a Famos auto sampler (LC Packings) onto the column. A gradient was formed, and peptides were eluted with increasing concentrations of solvent B (97.5% acetonitrile, 0.1% formic acid).

As peptides eluted, they were subjected to electrospray ionization and then entered into an LTQ Orbitrap Velos Pro ion-trap mass spectrometer (Thermo Fisher Scientific). Peptides were detected, isolated and fragmented to produce a tandem mass spectrum of specific fragment ions for each peptide. Peptide sequences (and hence protein identity) were determined by matching protein databases with the acquired fragmentation pattern by the software program Sequest v28 (rev. 13; Thermo Fisher Scientific)[39]. All databases include a reversed version of all the sequences, and the data were filtered to a peptide false-discovery rate of 1–2%. Only proteins that exhibited the following criteria were reported: at least five unique peptides detected in both PA1018 samples; and at least threefold enrichment in both PA1018 samples compared to the corresponding PAO1 and PA1009 samples.

微信公众号: 全球首发刊王

In vivo targeted pulldowns

Overnight cultures of *P. aeruginosa* strains encoding combinations of H–*Pa*LpxC, F–*Pa*MurA(WT) and F–*Pa*MurA(C117S) were diluted to OD_{600nm} = 0.01 in 50 ml LB and grown at 37 °C for 2.5 h. IPTG was added to each culture to a final concentration of 1 mM and cultures were grown for an additional 30 min at 37 °C before the addition of DMSO to 0.1% and CHIR-090 to 0.5 µg ml^{-1} where indicated. Cultures were incubated at 37 °C for an additional hour and collected by centrifugation at 12,000*g* for 10 min. Cells were washed in 1 ml LB and pelleted by centrifugation at 12,000*g* for 2 min. Cells were resuspended in 1 ml lysis buffer (25 mM Tris pH 8.0, 150 mM NaCl, 2% glycerol, 0.1% Triton X-100, 5 mM imidazole, 0.1% DMSO with or without 0.5 µg ml^{-1} CHIR-090 as indicated) and lysed by sonication on ice at 40% amplitude with four cycles of 15 s on, 15 s off. The extracts were clarified by two consecutive centrifugation steps at 21,130*g* for 5 min at 4 °C and protein concentration was determined with the Non-Interfering protein assay kit (G-biosciences 786-005). Supernatants were diluted to 3.5 mg ml^{-1} protein in 500 µl lysis buffer and mixed with 25 µl packed Ni-NTA resin (Qiagen 30230). Mixtures were incubated at 4 °C for 3 h rotating end-over-end. Beads were washed three times with 500 µl wash buffer (25 mM Tris pH 8.0, 150 mM NaCl, 2% glycerol, 0.1% Triton X-100, 25 mM imidazole, 0.1% DMSO with or without 0.5 µg ml^{-1} CHIR-090 as indicated). Protein was eluted in 40 µl elution buffer (25 mM Tris pH 8.0, 150 mM NaCl, 2% glycerol, 0.1% Triton X-100, 500 mM imidazole).

A 10 µg quantity of protein per input sample and 10 µl of eluate samples were resolved by SDS–PAGE in 10% polyacrylamide gels. Protein was transferred to PVDF membranes using the mixed molecular weight protocol on a Bio-Rad Trans-Blot Turbo transfer system and membranes were blocked in TBST + 5% milk at room temperature for 1 h. Membranes were then probed with anti-His (GenScript A00186-100) or anti-Flag (Sigma-Aldrich F7425) diluted 1:3,000 in TBST + 5% milk at room temperature for 1 h. Membranes were washed three times for 5 min each in TBST and probed with peroxidase-conjugated goat-anti-mouse (Rockland 610-1302) or rabbit TrueBlot: anti-rabbit IgG HRP (Rockland 18-8816-33) diluted 1:3,000 in TBST + 5% milk at room temperature for 1 h. Membranes were washed three times for 5 min each in TBST, developed with SuperSignal West Pico PLUS Chemiluminescent Substrate (Thermo 34580), and imaged using an Azure Biosystems C600 imager.

In vitro co-affinity purification

Purified F–*Pa*LpxC and H–MurA variants were each diluted to 2.5 µM in co-purification buffer (25 mM Tris pH 8.0, 150 mM NaCl, 10% glycerol, 1 mM dithiothreitol (DTT), 0.5% DMSO) to a final volume of 100 µl. When indicated, CHIR-090 was added to 5.7 µM. Mixtures were incubated on ice for 30 min after which 85 µl of each mixture was added to an equal volume of Anti-Flag M2 Magnetic Beads (Sigma-Aldrich M8823) and incubated at 4 °C for 1 h rotating end-over-end. Beads were collected with a magnetic rack and washed twice with 300 µl co-purification buffer and once with 200 µl co-purification buffer. When indicated, 5.7 µM CHIR-090 was maintained in the wash buffers. Proteins were eluted with 85 µl elution buffer (25 mM Tris pH 8.0, 150 mM NaCl, 10% glycerol, 1 mM DTT, 100 µg μl^{-1} Flag peptide (Millipore Sigma F3290)) by incubation at room temperature for 30 min rotating end-over-end. Input samples (5 µl) and eluate samples (15 µl) were resolved by SDS–PAGE in 4–20% polyacrylamide Mini-PROTEAN TGX gels (Bio-Rad 4561095) and protein was detected with Coomassie staining.

SEC

F–*Pa*LpxC and H–*Pa*MurA(WT) either individually or in combination were diluted to 7.5 µM each in 500 µl of SEC buffer 1 (10% glycerol, 0.33% DMSO, 1 mM DTT, 25 mM Tris pH 8.0 and 150 mM NaCl). When indicated, CHIR-090 was added to 37.5 µM. The mixtures were subsequently subjected to SEC using an ActaPure system equipped with a Superdex 200 10/300 GL column equilibrated with SEC buffer 2 (10% glycerol, 1 mM DTT, 25 mM Tris pH 8.0 and 150 mM NaCl) and eluate fractions were collected every 0.35 ml. For F–*Pa*LpxC and H–*Pa*MurA(WT) mixtures, three fractions corresponding to the shifted peak (14.1–15.2 ml with no CHIR-090, 13.8–14.8 ml for with CHIR-090) were pooled and concentrated by centrifugation at 4 °C using centrifugal filters with a molecular weight cutoff of 10-kDa (Amicon UFC801024). The concentrated protein was resubjected to SEC on a Superdex 200 10/300 GL column as described above. A 15 µl volume of the relevant fractions was resolved on a 4–20% polyacrylamide Mini-PROTEAN TGX gels (Bio-Rad 4568094) and protein was detected with Coomassie staining.

Growth curves

Cultures grown overnight at 30 °C were pelleted at 12,000*g* for 2 min and resuspended in fresh LB. Cultures were diluted in 96-well microtitre plates to an OD_{600nm} of 0.01 in 200 µl LB. Cultures were grown in a Tecan Infinite M-Plex plate reader at 37 °C and OD_{600nm} was monitored every 4 min with shaking at a 1-mm orbital for 200 s after each time point.

LPS silver stain

Overnight cultures were diluted to an OD_{600nm} of 0.01 in 25 ml LB cultures and grown for 4 h at 37 °C. For MurA- and MurB-depletion strains grown in the presence of inducer, 1 mM IPTG was present during the entire outgrowth. For MurA and LpxC overexpression strains, 1 mM IPTG or 0.1% arabinose was added after 3 h of outgrowth and incubation was continued for one additional hour. A 20 ml volume of each culture was centrifuged at 12,000*g* for 5 min and the pellet was washed in 1 ml of LB before centrifugation at 12,000*g* for 2 min. Pellets were resuspended to an OD_{600nm} of 20 in LDS sample buffer (Invitrogen NP0008) + 4% β-mercaptoethanol and boiled at 100 °C for 10 min. Samples were sonicated with a Q125 Qsonica sonicator at 25% amplitude for 1 s on, 1 s off for 30 cycles and protein concentration was determined using the Non-Interfering protein assay kit (G-biosciences 786-005). A 5 µg quantity of protein was loaded onto a 15% polyacrylamide–SDS gel and subjected to electrophoresis at 150 V for 70 min before western blot analysis of RpoA as described above. For the LPS gel, 50 µl of each sample was mixed with 1.25 µl of proteinase K (NEB P8107S) and incubated at 55 °C for 1 h followed by 95 °C for 10 min. The equivalent volume accounting for 10 µg of protein was resolved on a 4–12% Criterion XT Bis-Tris gel (Bio-Rad 3450124) at 100 V for 1 h 45 min and LPS was detected by silver stain as described previously[40]. Briefly, the gel was fixed by incubation in 200 ml of 40% ethanol, 5% acetic acid overnight. Periodic acid was added to 0.7% and allowed to incubate for 5 min. The gel was subsequently washed with three changes of 200 ml ultrapure water over the course of 2 h. The gel was impregnated with 150 ml of freshly made staining solution (0.018 N NaOH, 0.4% NH_4OH and 0.667% silver nitrate) for 10 min and subsequently washed with three changes of 200 ml ultrapure water over the course of 2 h. The gel was developed with 200 ml freshly prepared developer solution (0.26 mM citric acid pH 3.0, 0.014% formaldehyde) and development stopped in 100 ml of 1% acetic acid. The gel was imaged in a Bio-Rad ChemiDoc MP Imaging System.

Microscopy

Overnight cultures were diluted to an OD_{600nm} of 0.01 in 3 ml LB with 1 mM IPTG as appropriate. Cultures were grown at 37 °C for 2.5 h and, when appropriate, antibiotics were added to the following concentration: 0.5 µg ml^{-1} CHIR-090, 64 µg ml^{-1} fosfomycin or 0.5 µg ml^{-1} CHIR-090 + 16 µg ml^{-1} fosfomycin. Cultures were grown at 37 °C for an additional 1.5 h and 1 ml was pelleted at 12,000*g* for 2 min. Cell pellets were resuspended to an OD_{600nm} of 10 in LB, spotted onto an LB + 1.5% agar pad, and imaged on a Nikon Eclipse 50i microscope equipped with a 100× objective and a Photometrics CoolSNAP HQ^2 camera. Images were uniformly edited using the Fiji image analysis platform version 2.3.0/1.53q[41]. Cell width was quantified using the MicrobeJ plugin (version 5.13l) of ImageJ using the following settings: area: 250–max, length: 3–100, width: 5–25, width variation: 0–0.25, width symmetry: 0–1, circularity, 0.01–1. Normality of each population was tested and determined to be lacking by both the D'Agostino and Pearson test and the Anderson–Darling test carried out by Prism (GraphPad Software, LLC). Statistical significance was determined by Kruskal–Wallis test carried out by Prism (GraphPad Software, LLC). Microscopy source images are available at https://doi.org/10.5281/zenodo.7455522 (ref. [42]).

Determination of LpxC product and substrate levels in vivo

Overnight cultures were diluted to an OD_{600nm} of 0.01 in 200 ml LB, grown at 37 °C for 3 h, and IPTG was added to a final concentration of 1 mM. After incubation at 37 °C for one additional hour, cells were collected by centrifugation at 12,000*g* for 5 min. After resuspension in LB, cells were re-pelleted by centrifugation at 12,000*g* for 2 min. Pellets were then resuspended in 400 µl HPLC-grade H_2O. Subsequently, 250 µl chloroform and 500 µl methanol were added to the mixture. Samples were vortexed vigorously, placed on ice for 10 min, and then spun down at 4 °C at 21,000*g*. A 500 µl volume of the aqueous layer was collected and dried down at 30 °C overnight in a rotary evaporator. Dried material was resuspended in 100 µl ultrapure H_2O and the *Pa*LpxC product and substrate were detected by liquid chromatography–tandem mass spectrometry (LC–MS/MS) as described below.

In vitro LpxC activity assay

Purified F–*Pa*LpxC and H–MurA variants were diluted to 200 nM in binding buffer (25 mM Tris pH 8.0, 150 mM NaCl, 6% glycerol) to a final volume of 15 µl and mixtures were allowed to equilibrate to 30 °C for 10 min. The reaction was started by the addition of 15 µl of substrate mixture (25 mM Tris pH 8.0, 150 mM NaCl, 4% DMSO, 200 µM UDP-3-*O*-(*R*-3-hydroxydecanoyl)-GlcNAc (Carbosynth MU75071)) and allowed to proceed for 5 or 10 min at 30 °C at which point the reaction was stopped by the addition of 45 µl 2% acetic acid. Stopped reactions were frozen on dry ice, lyophilized, and the lyophilized material was resuspended in 60 µl H_2O. *Pa*LpxC product and substrate were detected by LC–MS as described below. The linear range of the reaction was determined as described above with purified F–*Pa*LpxC alone and reactions were stopped after 1, 2, 5, 10, 15 or 20 min (Extended Data Fig. 5b). H–*Pa*MurA(WT) and H–*Pa*MurA(C117S) alone exhibited background levels of LpxC activity, confirming that the observed stimulation of LpxC activity was not due to contamination within the protein preparation or an alternative enzymatic activity of *Pa*MurA (Extended Data Fig. 6e).

Activity assays with the *L. pneumophila* and *E. coli* LpxC–MurA pairs were carried out as described above except that concentration of H–MurA variants in the final reaction mixture was 200 nM. *L. pneumophila* reactions were stopped after 10 min and *E. coli* reactions were stopped after 1 min.

LC–MS

To quantify the activity of LpxC in vitro, the *Pa*LpxC product and substrate were detected by LC–MS on an Agilent Technologies 1200 series HPLC system in line with an Agilent 6520 Q-TOF mass spectrometer with electrospray ionization–mass spectrometry operating in negative mode. Samples were separated on a Waters Symmetry C18 column (5 µm; 3.9 mm × 150 mm) with a matching column guard using the following method: flow rate = 0.5 ml min^{-1}, 95% solvent A (H_2O, 10 mM ammonium formate) and 5% solvent B (acetonitrile) for 5 min

followed by a linear gradient of solvent B from 5–80% over 20 min. Agilent MassHunter Workstation Qualitative Analysis software version B.06.00 was used for analysing the MS data. The *Pa*LpxC substrate (UDP-3-*O*-(*R*-3-hydroxydecanoyl)-*N*-acetylglucosamine) was quantified by monitoring the abundance of 776.21 *m*/*z* and resolved as a single peak, which was integrated to infer substrate concentration. The *Pa*LpxC product (UDP-3-*O*-(*R*-3-hydroxydecanoyl)-glucosamine) was quantified by monitoring the abundance of 734.1944 *m*/*z* and often resolved as multiple peaks as reported previously[43], all of which were integrated to infer product concentration.

*Pa*LpxC substrate and product from the aqueous fraction of methanol–chloroform-extracted whole-cell lysates were analysed by LC–MS/MS using the same settings as the LC–MS analysis described above with the following adaptations. Parent ions with *m*/*z* of 776.1986 (corresponding to the *Pa*LpxC substrate) and *m*/*z* of 734.1872 (corresponding to the *Pa*LpxC product) were targeted for MS/MS with a collision energy of 40. Fragment ions between 50 and 850 *m*/*z* were analysed. The relative abundance of the *Pa*LpxC substrate and product were quantified by integrating the peaks observed for 776.1986 *m*/*z* and 734.1872 *m*/*z*, respectively. Raw LC–MS/MS data are available at https://doi.org/10.5281/zenodo.7455522 (ref. [42]).

In vitro MurA activity assay

Purified MurA variants were diluted to 200 nM (for *Pa*MurA(WT), *Pa*MurA(G58D) and *Pa*MurA(E406K)) or 2 µM (for *Pa*MurA* variants) in MurA reaction buffer (25 mM Tris pH 7.75, 6% glycerol) to a final volume of 25 µl and allowed to equilibrate to room temperature for 30 min. The reaction was started by the addition of 25 µl of MurA substrate mixture (25 mM Tris pH 7.75, 2 mM UDP-*N*-acetylglucosamine (Sigma-Aldrich U4375), and 1 mM phosphoenolpyruvate (Sigma-Aldrich 10108294001)) and the reaction was allowed to proceed for 30 min at 37 °C. The reaction was stopped by the addition of 800 µl Lanzetta reagent (0.033% malachite green, 1% ammonium molybdate, 1 N HCl, 0.2% Triton X-100)[44] and was incubated at room temperature for 1.25 min before the addition of 100 µl 34% sodium citrate. After an additional 30-min incubation at room temperature, the absorbance at 600 nm (A_{660nm}) of each sample was determined. A standard curve made from NaH_2PO_4 diluted in ultrapure water was used to quantify the amount of free P_i released by the reaction.

MurA–LpxC complex structure prediction

The structure of the LpxC–MurA complex was predicted using the default parameters of AlphaFold[19]. The *P. aeruginosa* PAO1 LpxC and MurA sequences or the *E. coli* MG1655 LpxC and MurA sequences were separated by a linker containing 15 repeats of Gly-Gly-Gly-Ser, which was not displayed for clarity. The structure was visualized using ChimeraX version 1.1.1 (ref. [45]).

Predicting complex interactions

We sought to use evolutionary covariation to predict the organisms in which the LpxC–MurA interaction does or does not exist. We used EVComplex run with EVcouplings v0.0.5 (ref. [18]) informed by the AlphaFold model to define an interaction score based on the inferred pairwise interaction terms J_{ij} of the complex and validate our proposed complex structure.

We first generated multiple sequence alignments of LpxC and MurA using jackhammer v3.1b2 (ref. [46]) with five iterations on the Uniref100 dataset[47] released in February 2021. Both *P. aeruginosa* and *E. coli* MurA (Uniprot: Q9HVW7 and P0A749) and LpxC (Uniprot: P47205 and P0A725) were used as initial sequences. In each case, alignments were constructed with a range of sequence score thresholds from 0.1 to 0.8 bits per residue, which did not meaningfully change which sequences were included for either species or protein. The same sequences were collected when starting from either species. For all analyses, we began with alignments of 54,940 MurA and 23,634 LpxC sequences generated from a relative bitscore of 0.8, and then concatenated MurA and LpxC sequences for each species, resolving ambiguities in pairing by selecting the sequence in each species closest to the original *P. aeruginosa* homologue. The final concatenated alignment contained 11,834 sequences. From this concatenated multiple sequence alignment, evolutionary couplings were determined using pseudo-likelihood maximization[48].

We more closely analysed the top five ranked intermolecular couplings (J_{ij}) of the direct coupling analysis model that correspond to pairs of amino acid positions in structural contact in the AlphaFold2 model, defined as having Cα atoms within 9 Å. Each pair of sequences can be scored on the coupling between MurA position i and LpxC position j on the basis of the appropriate J_{ij} term from the model. Summing these five scores for each species, we observed a bimodal distribution (Extended Data Fig. 9c). We selected only pairs of sequences whose sum over the five scores exceeded 0.1, totalling 1,689 sequence pairs (Extended Data Fig. 9c), and trained a new EVcomplex model on just these sequences from the concatenated alignment. In this model, five of the top six intermolecular couplings corresponded to predicted contacts in the AlphaFold2 structure (Extended Data Fig. 9d). We defined a final ‘interaction score’ based on the top 20 coupling terms from this refined model and used it to score all 11,834 pairs of sequences in the concatenated alignment again. The sequence analysis files are available at https://doi.org/10.5281/zenodo.7455522 (ref. [42]). The top 20 terms were chosen arbitrarily, but this exact choice does not meaningfully affect downstream analyses, and all conclusions about relative interaction scores remain true for choices of top couplings between five and five hundred.

To estimate the percentage of interacting sequences in each clade, we fitted a two-component Gaussian mixture model using sklearn v1.0.2 (ref. [49]) to the interaction scores for all sequence pairs and calculated the posterior probability of each sequence being drawn from the upper distribution. We mapped the taxonomic identifier for each sequence pair to its class in the National Center for Biotechnology Information taxonomy database using ETE3 v1.3.2 (ref. [50]), identifying 2,459 alphaproteobacteria and 2,781 gammaproteobacteria in the alignment. We then estimated the expected fraction of interacting sequences in each clade as the average of the interaction probabilities of each sequence in that clade.

To visualize the phylogenetic structure of these sequences, we built a maximum-likelihood phylogeny of MurA and LpxC from the concatenated multiple sequence alignment with FastTree version 2.1.11 (refs. [51]). Interaction scores were mapped onto the tree in Python v3.8.8 and visualized with the interactive tree of life (ITOL) v5 (ref. [52]). Custom code used in this study can be found at https://github.com/samberry19/evcomplex-interaction-scoring or https://doi.org/10.5281/zenodo.7471436 (ref. [53]).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

LC–MS/MS source data, microscopy images and sequence analysis files used to derive LpxC–MurA interaction scores are available at https://doi.org/10.5281/zenodo.7455522. Uniprot accession codes for genes used to generate LpxC–MurA interaction scores are included in the Methods and source data. All bacterial strains and plasmids developed in this study are available upon request. Source data are provided with this paper.

Code availability

All code generated in this study is available at https://github.com/samberry19/evcomplex-interaction-scoring or https://doi.org/10.5281/zenodo.7471436.

35. Miller, J. *Experiments in Molecular Genetics* (Cold Spring Harbor Laboratory, 1972).
36. Yang, D. C., Tan, K., Joachimiak, A. & Bernhardt, T. G. A conformational switch controls cell wall-remodelling enzymes required for bacterial cell division. *Mol. Microbiol.* **85**, 768–781 (2012).
37. Shevchenko, A., Wilm, M., Vorm, O. & Mann, M. Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels. *Anal. Chem.* **68**, 850–858 (1996).
38. Peng, J. & Gygi, S. P. Proteomics: the move to mixtures. *J. Mass Spectrom.* **36**, 1083–1091 (2001).
39. Eng, J. K., McCormac, A. L. & Yates, J. R. III An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
40. Tsai, C.-M. & Frasch, C. E. A sensitive silver stain for detecting lipopolysaccharide in polyacrylamide gels. *Anal. Biochem.* **119**, 115–119 (1982).
41. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
42. Hummels, K. R. et al. Data for "Coordination of bacterial cell wall and outer membrane biosynthesis". Zenodo https://doi.org/10.5281/zenodo.7455522 (2022).
43. Cohen, F. et al. Optimization of LpxC inhibitors for antibacterial activity and cardiovascular safety. *ChemMedChem* **14**, 1560–1572 (2019).
44. Lanzetta, P. A., Alvarez, L. J., Reinach, P. S. & Candia, O. A. An improved assay for nanomole amounts of inorganic phosphate. *Anal. Biochem.* **100**, 95–97 (1979).
45. Pettersen, E. F. et al. UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).
46. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinform.* **11**, 431 (2010).
47. Suzek, B. E. et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
48. Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: using pseudolikelihoods to inver Potts models. *Phys. Rev. E* **87**, 012707 (2013).
49. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
50. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
51. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
52. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
53. Berry, S. P. samberry19/evcomplex-interaction-scoring: v1.0 (v1.0). Zenodo https://doi.org/10.5281/zenodo.7471436 (2022).

Acknowledgements We thank all of the members of the Bernhardt, Rudner and Mekalanos labs for their thoughtful discussions and advice throughout this project, especially L. Marmont for technical insight regarding protein purification. We also thank S. Lory for strains and reagents as well as advice on the genetic manipulation of *P. aeruginosa*, B. Silva and M. Welsh for assistance with LC–MS/MS analysis, and the Taplin Mass Spectrometry Core Facility at Harvard Medical School for the identification of co-affinity-purified proteins. LC–MS data were acquired on an Agilent 6520 Q-TOF spectrometer supported by the Taplin Funds for Discovery Program. Protein structures were visualized with UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases. This work was supported by the National Institutes of Health (F32AI164630 to K.R.H., AI148752 to S.W., AI083365 to T.G.B., and U19 AI158028 to T.G.B and S.W.), the Chan Zuckerberg Initiative (CZI2018-191853 to D.S.M.) and Investigator funds from the Howard Hughes Medical Institute (T.G.B).

Author contributions K.R.H. and T.G.B. designed the study and wrote the manuscript. S.P.B., J.K.M. and D.S.M. devised and carried out the covariation analysis. Z.L. established the methanol–chloroform metabolite extraction protocol. A.T. and S.W. developed the LC–MS protocol. K.R.H. carried out all other experiments and data analysis. All authors read and approved the manuscript.

Competing interests The authors declare no competing interests.

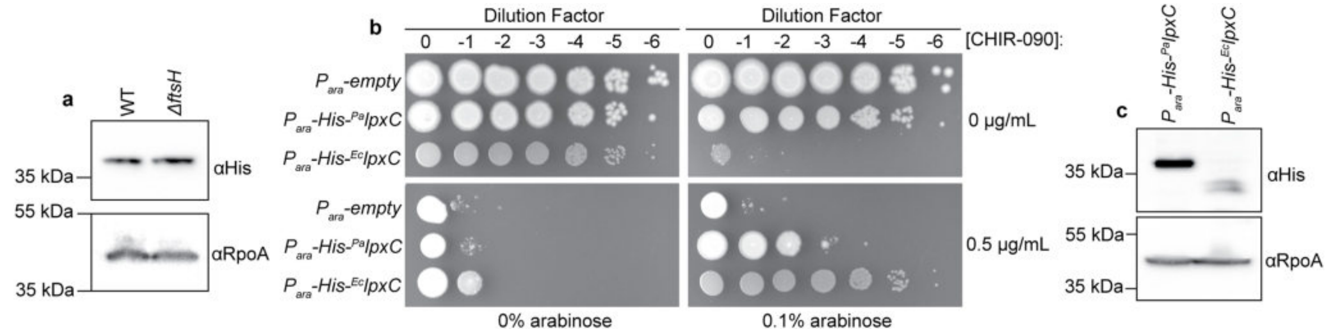
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-05750-0.

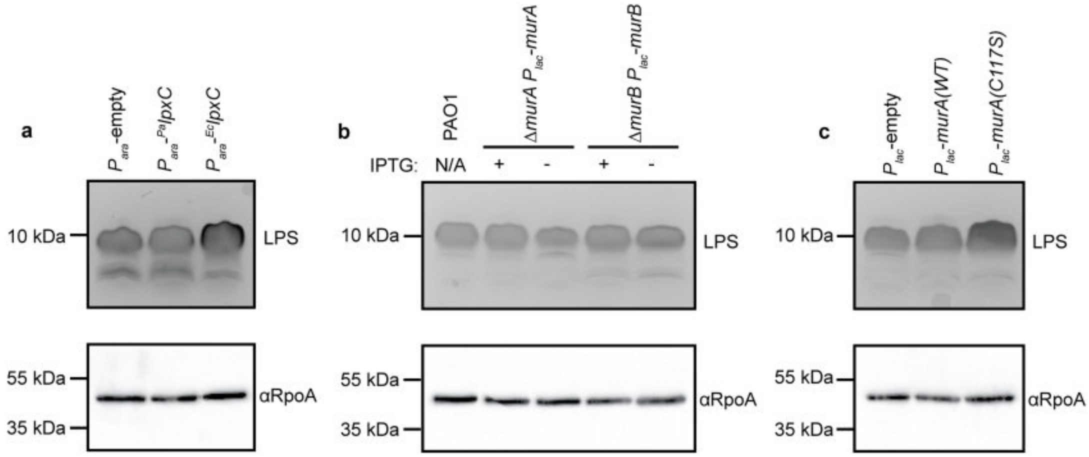
Correspondence and requests for materials should be addressed to Thomas G. Bernhardt.

Peer review information *Nature* thanks Russell Bishop and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

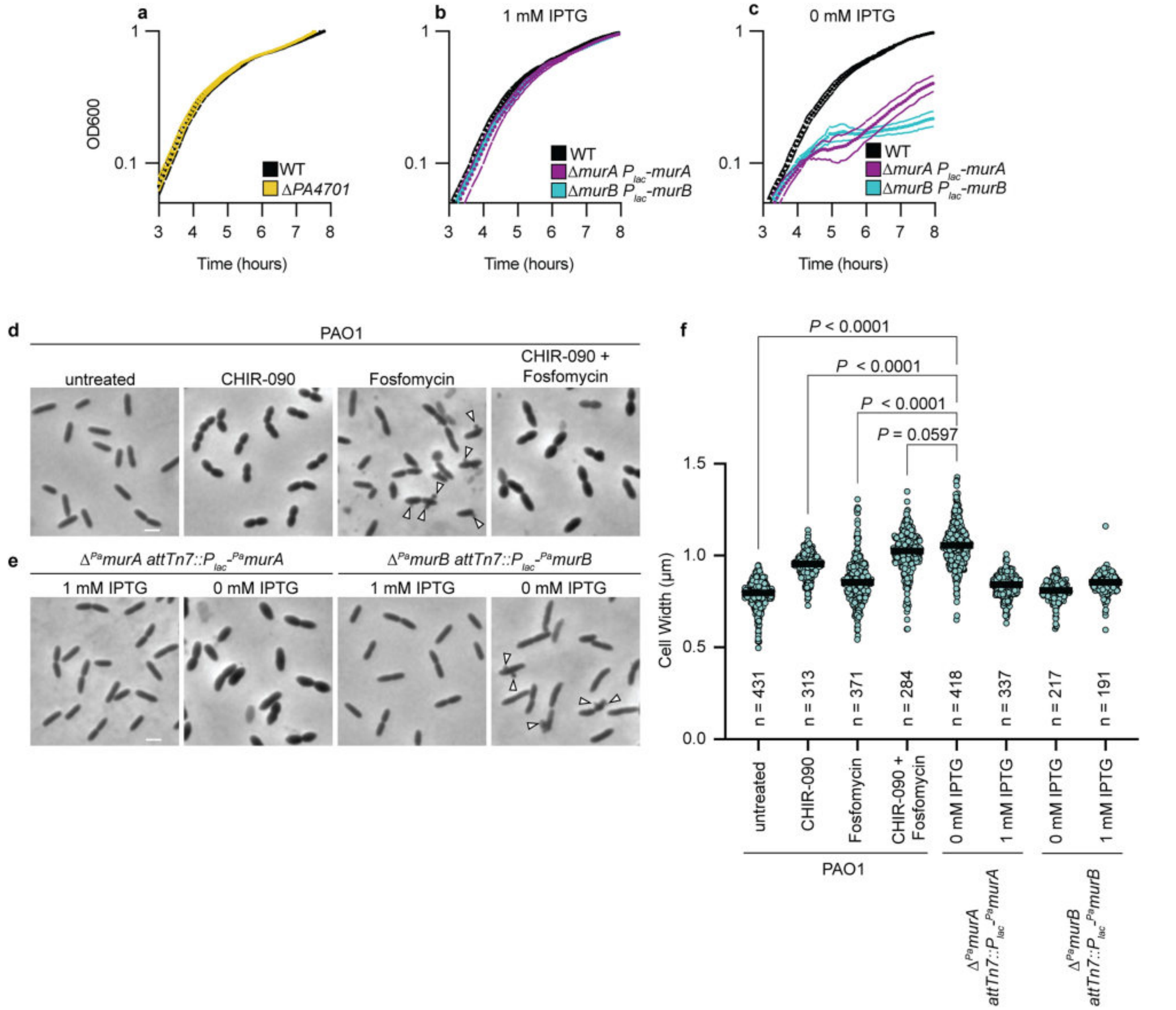
Reprints and permissions information is available at http://www.nature.com/reprints.



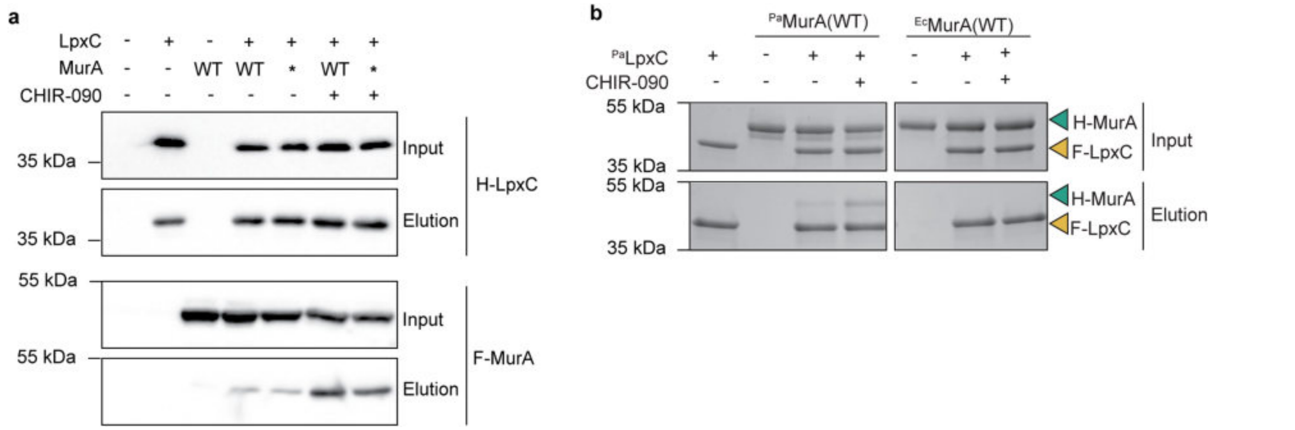
Extended Data Fig. 1 | Regulation of LpxC in *P. aeruginosa* differs from that observed in *E. coli*. (a) Anti-His immunoblot detecting H-^{Pa}LpxC expressed from the native chromosomal locus in wild-type cells or an *ftsH* deletion mutant. A corresponding blot for RpoA was used as a loading control. Data are representative of 3 biological replicates. (b) Spot titer assay in which serial dilutions of PAO1 harboring an empty plasmid or one encoding *His-^{Pa}lpxC* or *His-^{Ec}lpxC* under arabinose-inducible control were plated on LB agar supplemented with arabinose and/or the LpxC inhibitor CHIR-090 as indicated. Plates were incubated at 37 °C for 20 h before being imaged. Data are representative of 3 biological replicates. (c) Anti-His immunoblot analysis of His-^{Pa}LpxC or His-^{Ec}LpxC protein levels in exponentially growing PAO1. Immunoblot for RpoA serves as a loading control. Data are representative of 3 biological replicates. For gel source data, see Supplementary Fig. 1.



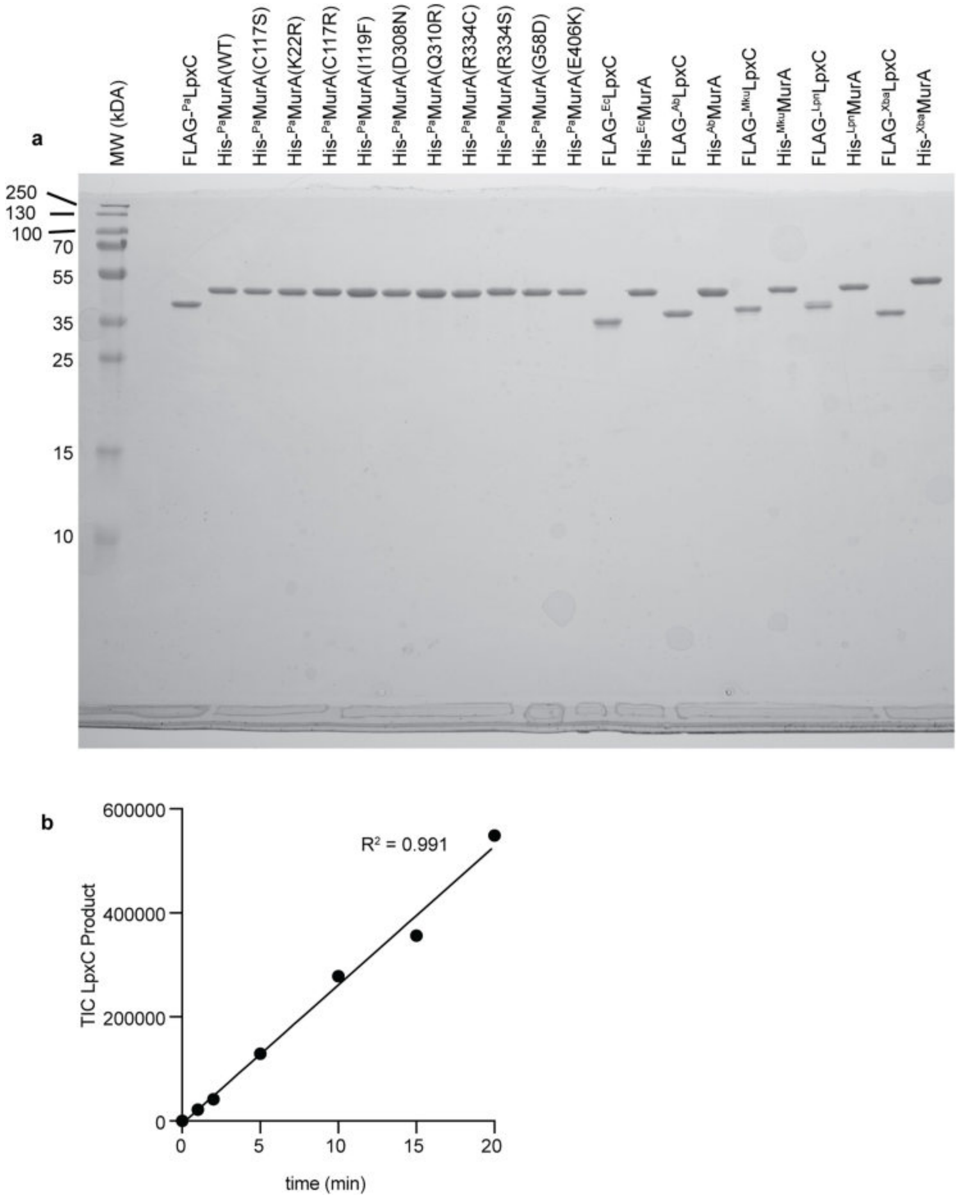
Extended Data Fig. 2 | LPS levels are altered upon mis-regulation of LpxC. Silver stain of LPS harvested from exponentially growing cultures and western blot of RpoA from the same samples as a loading control. (a) PAO1 harboring an empty plasmid or one encoding ^{Pa}*lpxC* or ^{Ec}*lpxC* under arabinose-inducible control were induced with 0.1% arabinose 1 h prior to harvesting samples. (b) PAO1, PA1118 [*ΔmurA* P_{lac}-*murA*], or PA1135 [*ΔmurB* P_{lac}-*murB*] were grown in the presence or absence of 1 mM IPTG as indicated before samples were processed. (C) PAO1 harboring an empty plasmid or one encoding ^{Pa}*murA(WT)* or ^{Pa}*murA(C117S)* under IPTG-inducible control were induced with 1 mM IPTG 1 h prior to harvesting samples. Data are representative of 3 biological replicates. For gel source data, see Supplementary Fig. 1.



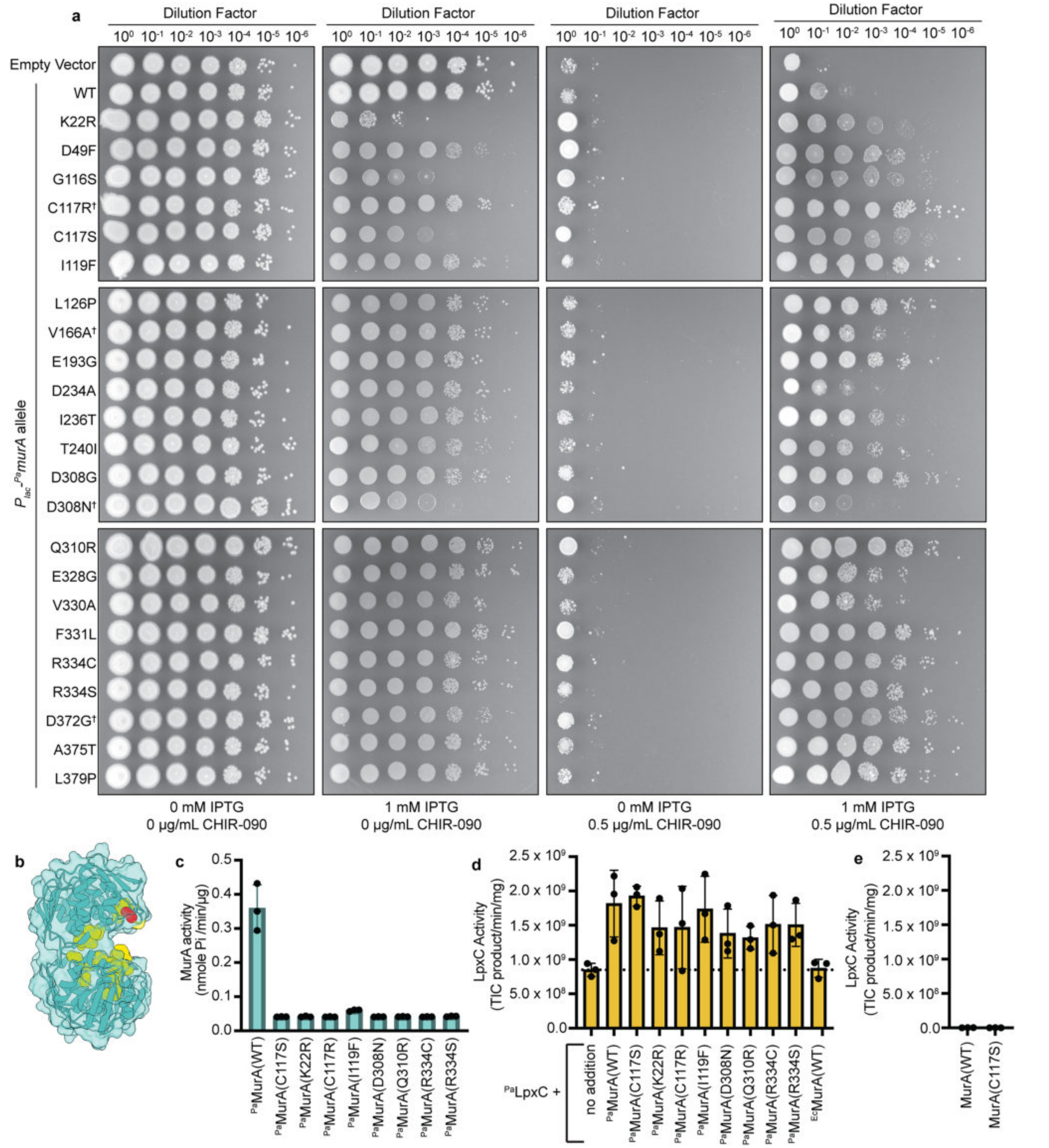
Extended Data Fig. 3 | MurA is essential and its depletion phenocopies simultaneous inhibition of PG and LPS biosynthesis. (a–c) Growth curves of *P. aeruginosa* strains in LB with or without IPTG as indicated. Dots represent the average of 3 biological replicates and dashed lines indicate the standard deviation. The following strains were used: (a) PAO1 [WT] and PA1080 [Δ*PA4701*], (b,c) PAO1 [WT], PA1118 [*ΔmurA P_{lac}-murA*], and PA1135 [*ΔmurB P_{lac}-murB*]. (d) Phase contrast images of *P. aeruginosa* cells after 1 hr treatment with the indicated antibiotic(s). Scale bar indicates 2 µm. Data are representative of at least 2 biological replicates (e) Phase contrast images of the indicated *P. aeruginosa murA* or *murB* depletion strains grown for 4 h in the presence or absence of inducer as indicated. MurB depletion was analyzed as a control to compare the phenotype of inactivating another early step in PG synthesis with that of MurA. Scale bar indicates 2 µm. Data are representative of at least 3 biological replicates (f) Quantification of cell width after 1 hr treatment with the indicated antibiotic(s) or after depletion of MurA or MurB. Each dot represents an individual cell and the median of the population is indicated by a black line. n indicates the number of cells analyzed. For each condition, the cells quantified were derived from a single population and data are representative of biological duplicates.



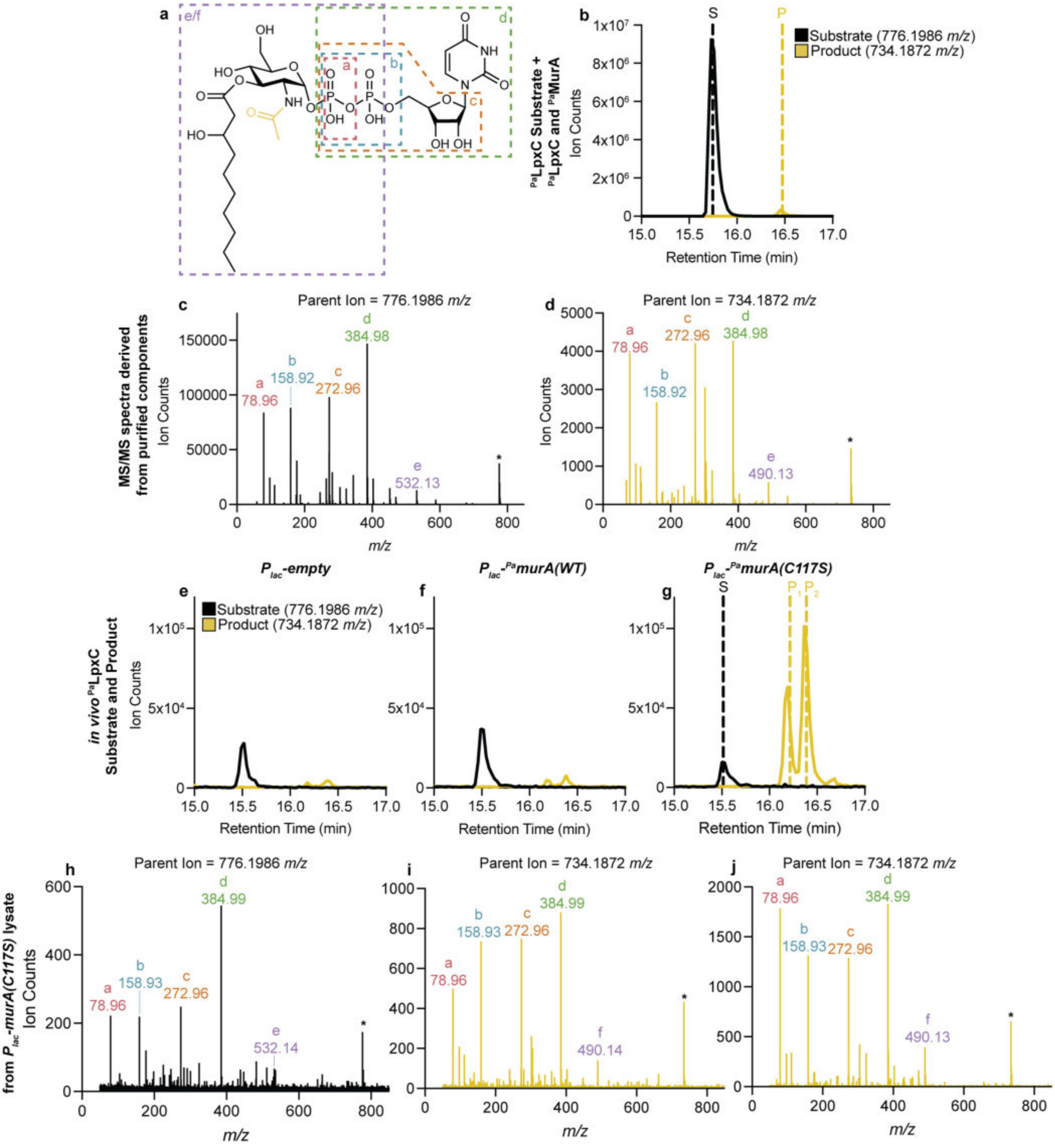
Extended Data Fig. 4 | *Pa*MurA interacts with *Pa*LpxC. (a) H-PaLpxC *in vivo* pulldowns. The expression status of H-PaLpxC and F-*Pa*MurA before (input) and after (elution) co-affinity purification using Ni-NTA resin is indicated above the immunoblots. The variant of F-*Pa*MurA produced is indicated by WT for F-*Pa*MurA(WT) or * for F-*Pa*MurA(C117S). When indicated, 0.5 µg/mL CHIR-090 was added to cultures 1 hr prior to harvesting and was maintained in all lysis and wash buffers as detailed in the methods section. The following strains were used to generate the lysates: PA239 (lane 1), PA1013 (lane 2), PA1068 (lane 3), PA1071 (lanes 4 and 6), and PA1121 (lanes 5 and 7). Data are representative of at 3 biological replicates. (b) Purified F-*Pa*LpxC (2.5 µM) was mixed in a 1:1 ratio with purified H-MurA variants in the presence or absence of CHIR-090 (5.7 µM) as indicated. The mixtures were pulled down with anti-FLAG resin and the input and elution subjected to SDS-PAGE and Coomassie staining. Mobilities of H-MurA and F-LpxC are indicated by a cyan or gold carrot, respectively. Data are representative of 3 replicates. For gel source data, see Supplementary Fig. 1.



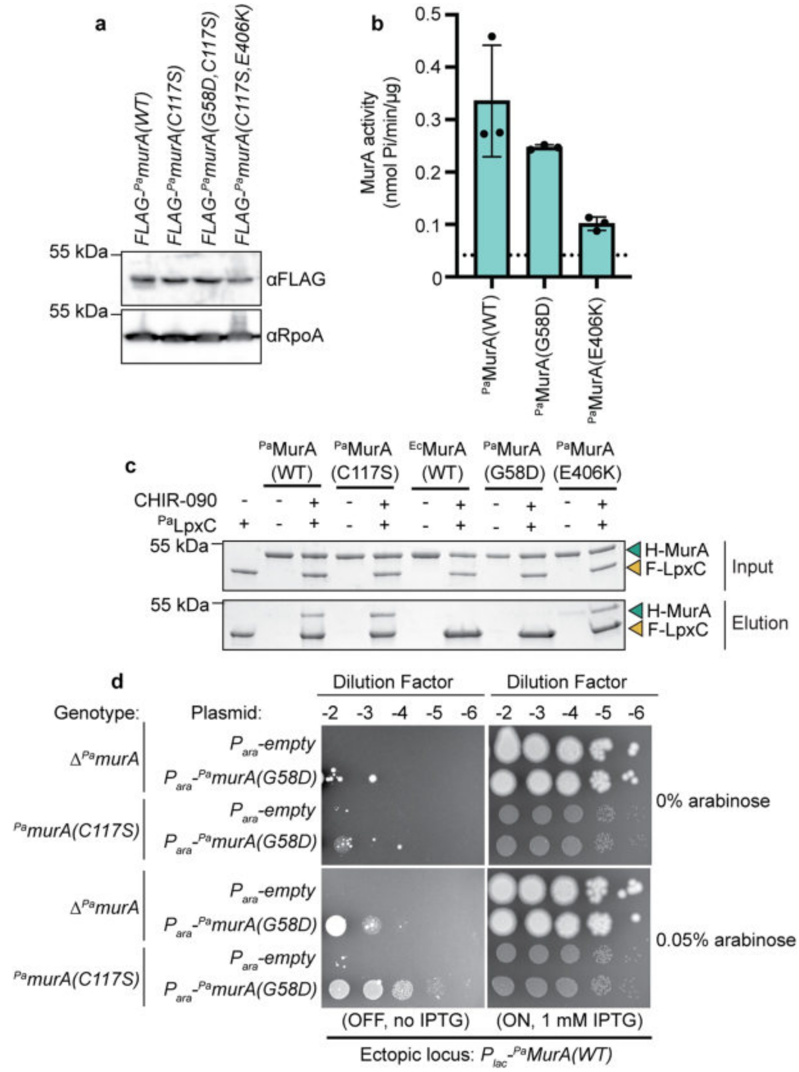
Extended Data Fig. 5 | Purified proteins used in this study and linearity of LpxC activity assay. (a) Purified proteins used in this study were resolved by SDS-PAGE and protein was visualized with Coomassie staining. Data are representative of at least two 2 replicates. For gel source data, see Supplementary Fig. 1. (b) Time course in which turnover of UDP-3-O-(R-3-hydroxydecanoyl)-N-acetylglucosamine (^{Pa}LpxC substrate) to UDP-3-O-(R-3-hydroxydecanoyl)-glucosamine (^{Pa}LpxC product) by FLAG-^{Pa}LpxC over the course of 20 min was monitored by LC-MS. The R^2 value of the linear regression is presented.



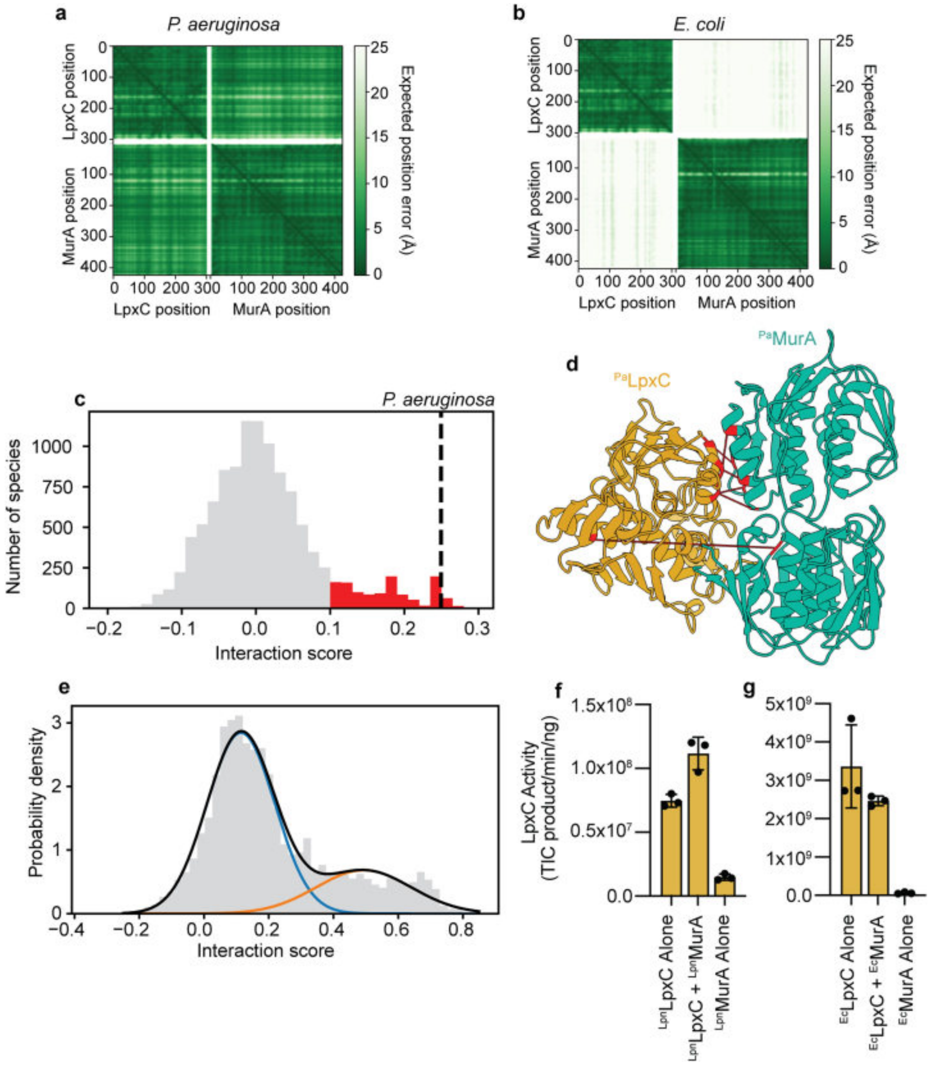
Extended Data Fig. 6 | Mutations in the ^{Pa}MurA active site are toxic but can be suppressed by inhibition of ^{Pa}LpxC. (a) Spot titer assay in which serial dilutions of PAO1 harboring an empty vector, one encoding ^{Pa}MurA(WT) or the indicated ^{Pa}MurA variant were plated on LB agar supplemented with IPTG and/or CHIR-090 as indicated. Plates were incubated at 37 °C for 20 h before imaging. † indicates the presence of a silent mutation in the construct. See Table S2 for details. Data are representative of 3 biological replicates. (b) Crystal structure of *E. cloacae* MurA (**PDB** 1EJC)[14] in which residues corresponding to the identified ^{Pa}MurA dominant negative alleles are depicted in gold spheres, or in the case of Cys117, a red sphere. Note that the substitutions all cluster around the active site. (c) MurA activity assay in which purified ^{Pa}MurA variants (100 nM) were mixed with UDP-GlcNAc (1 mM) and PEP (0.5 mM) and the release of P_i was measured by Lanzetta assay. Dots indicate the values obtained for three individual replicates, bars indicate the mean, and error bars represent their standard deviation. (d) Catalytic activity of purified ^{Pa}LpxC (100 nM) alone or in the presence of MurA variants (100 nM) assayed by conversion of UDP-3-O-(R-3-hydroxydecanoyl)-N-acetylglucosamine (^{Pa}LpxC substrate) to UDP-3-O-(R-3-hydroxydecanoyl)-glucosamine (^{Pa}LpxC product). Dots indicate the values obtained for three individual replicates, bars indicate the mean, and error bars represent their standard deviation. (e) LpxC enzymatic activity detected from preparation of MurA variants (100 nM) alone assayed as in panel d. Dots indicate the values obtained for three individual replicates, bars indicate the mean, and error bars represent their standard deviation.



Extended Data Fig. 7 | LC-MS/MS analysis of PaLpxC substrate and product. (a) Chemical structure of the PaLpxC substrate. The acetyl moiety removed by LpxC is highlighted in gold. Boxed regions indicate putative fragment ions highlighted in panels c, d, and h-j. (b) Extracted ion chromatogram (EIC) of UDP-3-O-(R-3-hydroxydecanoyl)-N-acetylglucosamine (PaLpxC substrate, 776.1986 *m/z*, black line) and UDP-3-O-(R-3-hydroxydecanoyl)-glucosamine (PaLpxC product, 734.1872 *m/z*, gold line) derived from *in vitro* reaction mixtures containing the PaLpxC substrate along with purified PaLpxC and PaMurA resolved using LC-MS operating in negative mode. The dashed lines indicate the peaks assigned to the PaLpxC substrate (S) and PaLpxC product (P). (c–d) MS/MS spectrum associated with the panel b peaks S and P, respectively. The parent ion is indicated by an asterisk and fragment ions corresponding to those highlighted in panel a are labeled. (e-g) EICs PaLpxC product and PaLpxC substrate detected by LC-MS in the aqueous fraction of methanol-chloroform extracted whole cell lysates. The dashed lines indicate the peak assigned to the PaLpxC substrate (S) and PaLpxC product peaks (P_1 and P_2). Note that the PaLpxC product has been previously reported to resolve as two peaks[43], both of which were integrated to infer the relative product abundance. Data are representative of biological triplicates. (h-j) MS/MS spectrum associated with the panel g peaks S, P_1, and P_2, respectively. The parent ion is indicated by an asterisk and fragment ions corresponding to those highlighted in panel a are labeled.



Extended Data Fig. 8 | MurA(G58D) and MurA(E406K) impact binding and activation of ^{Pa}LpxC. (a) Anti-FLAG immunoblot detecting F-^{Pa}MurA variants after 1 h of induction with 1 mM IPTG. A corresponding blot for RpoA was used as a loading control. Data are representative of 3 biological replicates. (b) MurA activity assay in which purified ^{Pa}MurA variants (100 nM) were mixed with UDP-GlcNAc (1 mM) and PEP (0.5 mM) and the release of P_i was measured by Lanzetta assay. Dots indicate the values obtained for three individual replicates, bars indicate the mean, and error bars represent their standard deviation. The dashed line indicates the average catalytic activity of ^{Pa}MurA(C117S) observed in Fig S7C. (c) *in vitro* pulldowns in which purified F-^{Pa}LpxC and H-MurA variants were mixed in a 1:1 ratio in the presence or absence of CHIR-090 and processed as in Extended Data Fig. 4b. Data are representative of at least two replicates. (d) Spot titer assay in which serial dilutions of a PAO1 strain harboring a ^{Pa}*murA* deletion or ^{Pa}*murA(C117S)* allele at the native locus complemented by a chromosomally-integrated, IPTG-inducible copy of ^{Pa}*murA(WT)* were plated on LB agar with the indicated supplements. As indicated, the strains also contained an empty plasmid or one encoding ^{Pa}*murA(G58D)* under arabinose-inducible control. Plates were incubated at 37°C for 20 h before being photographed. Data are representative of 3 biological replicates. For gel source data, see Supplementary Fig. 1.



Extended Data Fig. 9 | ^{Pa}LpxC and ^{Pa}MurA are predicted to interact. (a,b) AlphaFold2 predicted aligned error matrices of the ^{Pa}LpxC/^{Pa}MurA complex and ^{Ec}LpxC/^{Ec}MurA complex structures, respectively. (c) Distribution of initial interaction scores among all LpxC and MurA pairs analyzed. Red bars indicate the sequences used to train the final EVcomplex model and the score corresponding to the ^{Pa}LpxC/^{Pa}MurA pair is indicated by a dashed line. (d) Model structure of the ^{Pa}LpxC/^{Pa}MurA complex predicted by AlphaFold2[19]. ^{Pa}LpxC is represented in gold and ^{Pa}MurA is represented in cyan. The top six intermolecular couplings between LpxC/MurA residues from the final EVcomplex model are highlighted in red with red lines connecting the coupling pairs. (e) Distribution of final interaction scores among all LpxC and MurA pairs analyzed. The data fit a two-component Gaussian mixture model (black line) indicating the presence of a population with low interaction scores (blue line) and high interaction scores (orange line). (f) Catalytic activity of purified ^{Lpn}LpxC (100 nM) alone or in the presence of ^{Lpn}MurA (200 nM) assayed by conversion of UDP-3-O-(R-3-hydroxydecanoyl)-N-acetylglucosamine to UDP-3-O-(R-3-hydroxydecanoyl)-glucosamine. Dots indicate the values obtained for three individual replicates, bars indicate the mean, and error bars represent their standard deviation. (g) Catalytic activity of purified ^{Ec}LpxC (100 nM) alone or in the presence of ^{Ec}MurA (200 nM) assayed by conversion of UDP-3-O-(R-3-hydroxydecanoyl)-N-acetylglucosamine to UDP-3-O-(R-3-hydroxydecanoyl)-glucosamine. Dots indicate the values obtained for three individual replicates, bars indicate the mean, and error bars represent their standard deviation.

nature portfolio

Corresponding author(s): Thomas G. Bernhardt

Last updated by author(s): 12/22/2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Covariation analysis was performed using open-source softwares Jackhmmer v3.1b2, EVcouplings v0.0.5, Python v3.8.8, FastTree v2.1.11 and AlphaFold2.
Data analysis	Statistical analyses were performed by Prism 9 (GraphPad Software, LLC.) LC/MS data were analyzed using Agilent MassHunter Workstation Qualitative Analysis software version B.06.00. Protein sequences from pulldown experiments were determined using Sequest ver 28 (rev 13). Structural data were visualized using ChimeraX version 1.1.1 and phylogenetic data were visualized using the interactive tree of life version 5. Microscopy images were processed using ImageJ2 version 2.3.0/1.53q and the MicrobeJ version 5.13I plugin. The distribution LpxC/MurA interaction scores was analyzed using sklearn v1.0.2 and ETE3 v1.3.2.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data used to generate graphs presented in this work are available as source data. LC-MS/MS source data, microscopy images, and computational intermediates used to derive LpxC-MurA interaction scores can be accessed using the DOI 10.5281/zenodo.7455522. Uniprot accession codes for genes used to generate LpxC-MurA interaction scores can be found in the methods and Source Data 3. All bacterial strains and plasmids developed in this study are available upon request. All code generated in this study can be accessed https://github.com/samberry19/evcomplex-interaction-scoring or by using DOI 10.5281/zenodo.7471436.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. In the case of microscopy experiments, all experiments were performed using greater than 100 cells per population spanning multiple fields of view, which is standard for the field. All other experiments reported bulk measurements of population wide behaviors, involving greater than 1x10^8 cells (for in vivo assays) or greater than 1 x 10^12 protein copies (for in vitro assays) as is standard for the field and each experiment was performed at least twice.
Data exclusions	No data were excluded from this study
Replication	All data are from a minimum of two independent experiments. All replication attempts were successful.
Randomization	Samples were not randomized in this study. Covariates were controlled by reproducing the experiments on separate days.
Blinding	We did not blind samples as, after each experimental setup, all measurements and analyses were performed identically across all conditions.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Anti-His (GenScript A00186-100); Anti-FLAG (Sigma F7425); Anti-RpoA (BioLegend 663104); Goat-anti-Mouse (Rockland 610-1302); Rabbit TrueBlot®: Anti-Rabbit IgG HRP (Rockland 18-8816-33)

Validation

Anti-His and Anti-FLAG antibodies have been used numerous times to detect His and FLAG-tagged proteins by western blot and validation information can be found at the manufacturers' websites (https://www.genscript.com/antibody/A00186-THE_His_Tag_Antibody_mAb_Mouse.html and https://www.sigmaaldrich.com/US/en/product/sigma/f7425, respectively). In addition, we verified that neither antibody reacts with Pseudomonas aeruginosa lysates derived from strains that do not encode a His- or FLAG-tagged protein (see Extended Data 4A). Monoclonal Anti-RpoA Clone 4RA2 has been used extensively in the field to detect the RNA polymerase subunit alpha in diverse gram-negative bacteria and validation can be found at the manufacturer's website (https://www.biolegend.com/it-it/products/purified-anti-e-coli-rna-polymerase-alpha-antibody-14680). Further, this antibody has been shown to react with specifically with purified P. aeruginosa PAO1 RNA polymerase (Ceyssens, P. J. et al. The Phage-Encoded N-Acetyltransferase Rac Mediates Inactivation of Pseudomonas aeruginosa Transcription by Cleavage of the RNA Polymerase Alpha Subunit. Viruses 12, doi:10.3390/v12090976 (2020).)

Article

Autoimmunity in Down's syndrome via cytokines, CD4 T cells and CD11c$^+$ B cells

https://doi.org/10.1038/s41586-023-05736-y

Received: 16 December 2021

Accepted: 17 January 2023

Published online: 22 February 2023

Check for updates

Louise Malle[1,2,3,4,5], Roosheel S. Patel[1,2,3,4,5], Marta Martin-Fernandez[1,2,3,4,5,14], O'Jay Stewart[1,2,3,4,5,14], Quentin Philippot[6,7], Sofija Buta[1,2,3,4,5], Ashley Richardson[1,2,3,4,5], Vanessa Barcessat[4], Justin Taft[1,2,3,4,5], Paul Bastard[6,7,8,9], Julie Samuels[3], Clotilde Mircher[10], Anne-Sophie Rebillat[10], Louise Maillebouis[10], Marie Vilaire-Meunier[10], Kevin Tuballes[4], Brad R. Rosenberg[5], Rebecca Trachtman[2,3], Jean-Laurent Casanova[6,7,8,11,12], Luigi D. Notarangelo[13], Sacha Gnjatic[4], Douglas Bush[2] & Dusan Bogunovic[1,2,3,4,5] ✉

Down's syndrome (DS) presents with a constellation of cardiac, neurocognitive and growth impairments. Individuals with DS are also prone to severe infections and autoimmunity including thyroiditis, type 1 diabetes, coeliac disease and alopecia areata[1,2]. Here, to investigate the mechanisms underlying autoimmune susceptibility, we mapped the soluble and cellular immune landscape of individuals with DS. We found a persistent elevation of up to 22 cytokines at steady state (at levels often exceeding those in patients with acute infection) and detected basal cellular activation: chronic IL-6 signalling in CD4 T cells and a high proportion of plasmablasts and CD11c$^+$Tbet^{high}CD21^{low} B cells (Tbet is also known as TBX21). This subset is known to be autoimmune-prone and displayed even greater autoreactive features in DS including receptors with fewer non-reference nucleotides and higher *IGHV4-34* utilization. In vitro, incubation of naive B cells in the plasma of individuals with DS or with IL-6-activated T cells resulted in increased plasmablast differentiation compared with control plasma or unstimulated T cells, respectively. Finally, we detected 365 auto-antibodies in the plasma of individuals with DS, which targeted the gastrointestinal tract, the pancreas, the thyroid, the central nervous system, and the immune system itself. Together, these data point to an autoimmunity-prone state in DS, in which a steady-state cytokinopathy, hyperactivated CD4 T cells and ongoing B cell activation all contribute to a breach in immune tolerance. Our findings also open therapeutic paths, as we demonstrate that T cell activation is resolved not only with broad immunosuppressants such as Jak inhibitors, but also with the more tailored approach of IL-6 inhibition.

First desc ribed by John Langdon Down in 1866, Down's syndrome (DS) or trisomy 21 is the most common chromosomal anomaly in the USA today, affecting 1 in 700 newborn babies[3,4]. This extra copy of around 200 genes results in a syndrome with considerable phenotypic variability that includes intellectual disability, developmental malformations—particularly of the heart and the gut—and increased risk of Alzheimer's disease[1]. As care for individuals with DS has substantially improved in recent decades[5], the immune features of DS have become apparent: patients have an increased risk of severe infectious disease concomitant with a higher incidence of autoimmunity including thyroiditis (50%), coeliac disease (5%), alopecia areata (1–11%) and type 1 diabetes (1%)[1,2,6].

Recent studies into the molecular mechanisms of immunological disease have focused on the overactive interferon (IFN) response[7–9] and thymic dysfunction[10] reported in individuals with DS as most IFN receptor subunits and *AIRE* are expressed from chromosome 21. On the innate side, monocytes from individuals with DS exhibit basal IFN-I and II signalling and hyper-respond to IFNα and IFNγ stimulation[7]. On the adaptive side, thymic architecture perturbations[10] and T cell polarization towards differentiated subsets have been described[11]. Furthermore, low B cell counts in individuals with DS have been documented for decades[12] and recent investigations have identified decreased proliferation and increased apoptosis in this population[13].

[1]Center for Inborn Errors of Immunity, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [2]Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [3]Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [4]Precision Immunology Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [5]Department of Microbiology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [6]Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM U1163, Necker Hospital for Sick Children, Paris, France. [7]University of Paris, Imagine Institute, Paris, France. [8]St Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, The Rockefeller University, New York, NY, USA. [9]Pediatric Hematology-Immunology and Rheumatology Unit, Necker Hospital for Sick Children, Assistance Publique-Hôpitaux de Paris (AP-HP), Paris, France. [10]Institut Jérôme Lejeune, Paris, France. [11]Department of Pediatrics, Necker Hospital for Sick Children, Paris, France. [12]Howard Hughes Medical Institute, New York, NY, USA. [13]Laboratory of Clinical Immunology and Microbiology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA. [14]These authors contributed equally: Marta Martin-Fernandez, O'Jay Stewart. ✉e-mail: Dusan.Bogunovic@mssm.edu

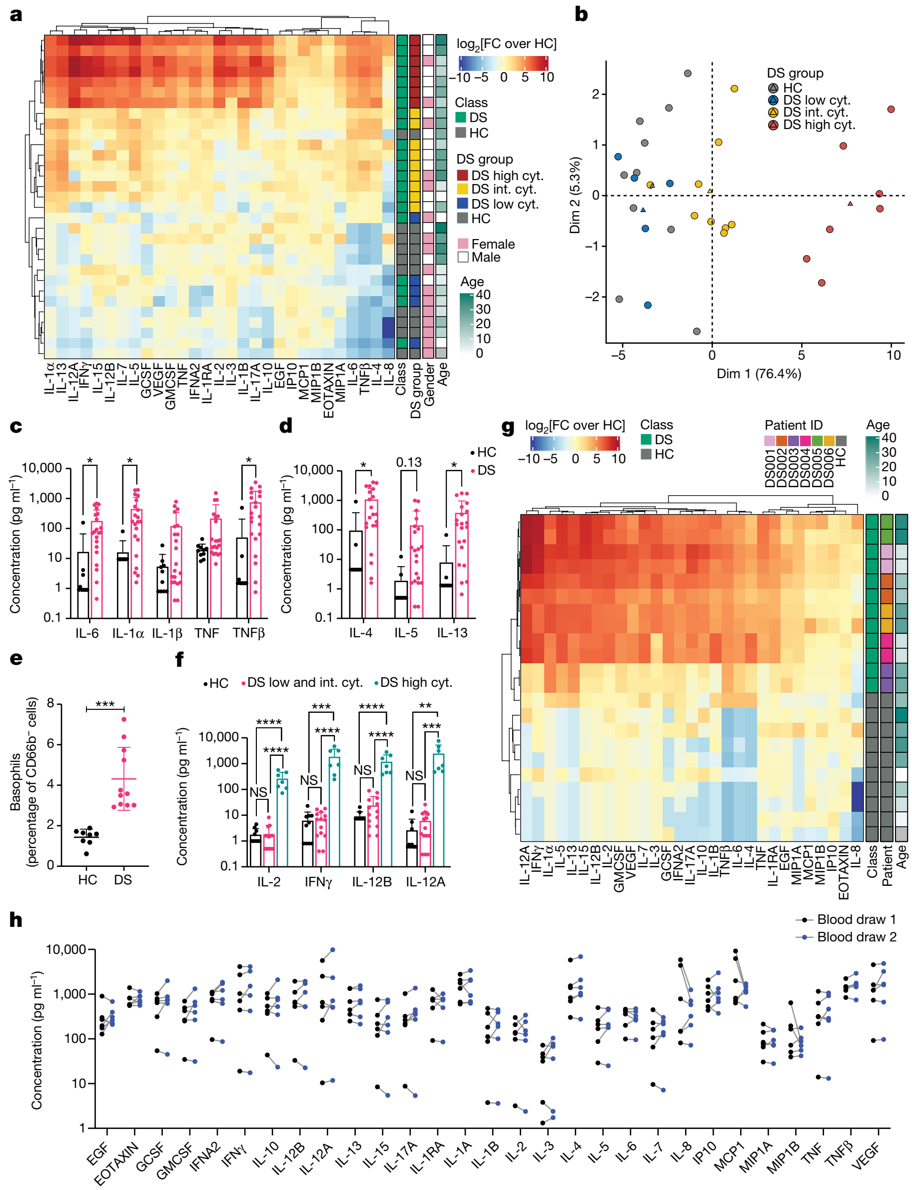


Fig. 1 | Cytokine profiling indicates broad immune dysregulation in most individuals with DS. a, Multiplex cytokine analysis using the magnetic Luminex assay of plasma from individuals with DS (*n* = 21) and HC individuals (*n* = 10), expressed as the log_2-transformed fold change (FC) over the mean HC per cytokine (cyt.). Unsupervised clustering of samples and cytokines using the complete method (distance metric, Euclidean). Int., intermediate. **b**, PCA analysis of serum cytokines from the samples. **c**,**d**, Raw values of acute phase proteins (**c**) and T_H2 cytokines (**d**) in the plasma of individuals with DS (*n* = 21) and HC individuals (*n* = 10) measured using the magnetic Luminex assay. **e**, The frequency of basophils expressed as the percentage of $CD66b^-$ cells (non-granulocytes) from adults with DS (*n* = 11) and age-matched HC individuals (*n* = 8). **f**, Raw values of IL-2 and T_H1 cytokines in the plasma of individuals with DS (*n* = 21) and HC individuals (*n* = 10) measured using the magnetic Luminex assay. **g**,**h**, Multiplex cytokine analysis using the magnetic Luminex assay of plasma from blood drawn at separate timepoints from individuals with DS (*n* = 6) expressed as log_2-transformed fold change over the mean HC individuals per cytokine followed by unsupervised clustering (**g**) and as raw values (**h**). No significant differences were detected on the basis of paired *t*-tests between blood draws for each individual for each cytokine ($P > 0.05$ for all pairs). For **c**–**f**, data are mean ± s.d. Significance was assessed using two-tailed unpaired *t*-tests (**c**–**e**) and ANOVA with Tukey's post hoc analysis (**f**); $*P \leq 0.05$; $**P \leq 0.005$; $***P \leq 0.0005$; $****P \leq 0.0001$.

It is now understood that a tightly regulated immune response is critical to prevent infection and excessive inflammation: the absence of a well-orchestrated immune response leads to opportunistic infections, whereas an overactive immune response leads to systemic organ damage[14]. Where DS lies on this spectrum of immune dysregulation and how it contributes to clinical manifestations is still largely unknown.

Down's syndrome is a cytokinopathy

To capture the soluble immune landscape at steady-state in DS, we performed a cytokine array on plasma from individuals with DS (*n* = 21) and age-matched healthy control (HC, *n* = 10) individuals. Donors had no signs of infection at sampling. On the basis of unsupervised hierarchical clustering on 29 analytes, individuals with DS segregated into 3 distinct categories (Fig. 1a,b). One-third had widespread soluble immune dysregulation with up to 2,000-fold elevation in 22 out of the 29 markers assayed. In 9 out of 21 individuals, a subset of cytokines was significantly elevated compared with in the control individuals (including IL-13, IL-4, TNFβ (also known as LTα) IL-6 and IL-1α). Finally, the remaining five individuals with DS clustered with HC individuals. As previously reported for HC individuals[15], there was a correlation between inflammatory cytokine profiles and age in DS (analysis of

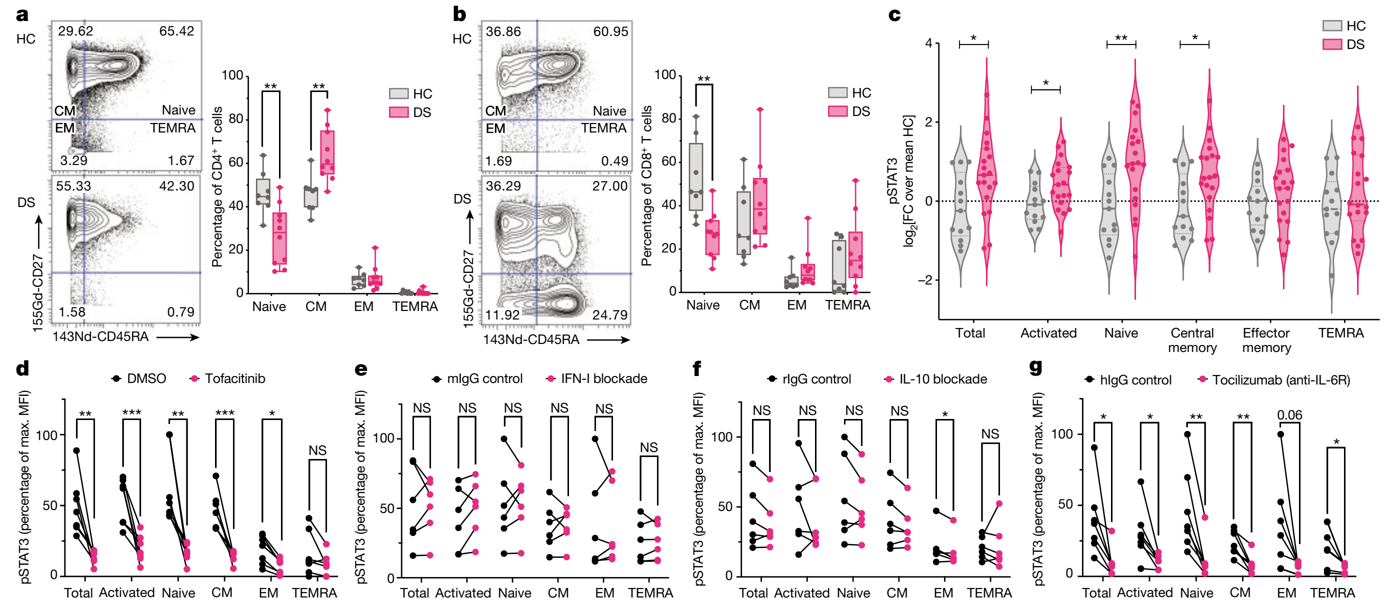


Fig. 2 | T cell activation in DS is rescued by Jak inhibition or IL-6 blockade. **a**,**b**, Representative plots and calculated frequencies of CD4$^+$ (**a**) and CD8$^+$ T cell naive, central memory (CM), effector memory (EM) and terminally differentiated effector memory (TEMRA) (**b**) subsets in whole blood from adults with DS (n = 10) and age-matched HC individuals (n = 8). **c**, Basal STAT3 phosphorylation in CD4$^+$ T cell subsets from individuals with DS (n = 19) and age-matched HC individuals (n = 13) and expressed as the $\log_2$-transformed fold change over the mean of HC individuals per subset. **d**–**g**, STAT3 phosphorylation in CD4$^+$ T cell subsets from individuals with DS, normalized to the maximum value per experiment after ex vivo whole-blood treatment for 4 h with JAK inhibition (tofacitinib) (500 nM) (n = 7) (**d**); IFN blockade (anti-IFNAR2 (5 μg ml^{-1}), anti-IFNα (0.2 μg ml^{-1}) and anti-IFNβ (0.2 μg ml^{-1}) antibodies) (n = 6) (**e**); IL-10 blockade (anti-IL-10 (5 μg ml^{-1}) and anti-IL-10R (5 μg ml^{-1}) antibodies) (n = 6) (**f**) or IL-6 blockade (tocilizumab, 50 μg ml^{-1}) (n = 7) (**g**). In **a** and **b**, the whiskers denote the minimum and maximum values, the box limits denote quartiles 1–3, and the centre bar denotes the mean. For **a**–**g**, significance was assessed using two-tailed unpaired t-tests; NS, not significant (P > 0.05).

variance (ANOVA), P = 0.0231; Extended Data Fig. 1a). After scoring samples based on clinical immunological manifestations (Methods and Extended Data Table 1), we observed a significant association between immune scores and the cytokine-based clusters (ANOVA, P = 0.0141; Extended Data Fig. 1b). It is not yet clear whether dysregulated cytokines drive clinical immune dysfunction in individuals with DS, or vice versa.

To determine the magnitude of this global cytokine dysregulation, we compared the cytokine profiles of individuals with DS and HC individuals at steady state, in mild or severe COVID-19 (n = 13 (control) and n = 7 (individuals with DS)) or another acute respiratory infection (n = 1 individual with DS) (Extended Data Fig. 1c). The COVID-19 samples were collected during hospitalization or at follow-up. Samples from donors with DS and with COVID-19 were available at only one point, whereas control samples were collected during the acute phase and at follow-up. Notably, the cytokine profiles of the patients with an infection (irrespective of ploidy, disease severity or infecting virus) were not significantly different from those of uninfected individuals with DS—unbiased hierarchical clustering placed patients with an infection across all three groups of individuals with DS. The uninfected high-cytokine DS group had a broader, more severe inflammatory profile compared with any patient with an infection. Our findings suggest that at least a third of individuals with DS have cytokine levels similar to severe acute infection at the baseline.

The acute-phase proteins IL-6, IL-1α and TNFβ were basally elevated in most individuals with DS (Fig. 1a,c), consistent with previous studies[16,17]. Intracellular staining suggested that CD16$^+$ monocytes, conventional dendritic cells (cDCs), and central memory CD4 and CD8 T cells from individuals with DS contained slightly more IL-6 (Extended Data Fig. 1d). T helper 2 (T_H2) cytokines IL-4 and IL-13, two central drivers of the allergic response[18], were significantly elevated in the plasma of individuals with DS (Fig. 1a,d), possibly explained by the concurrent increase in basophils (Fig. 1e). The role of other T cell subsets, especially T_H2 and T_H9 cells, must be further examined. IL-2, T_H1 cytokines IL-12 and IFNγ, and the T_H17 cytokine IL-17 were elevated in only the high-cytokine subgroup of individuals with DS (Fig. 1a,f).

Measuring cytokine levels in multiple blood samples, drawn 5 to 10 months apart, we found that the immune profile of individuals with DS was highly stable for both the high- and low-cytokine groups (Fig. 1g,h). Our findings suggest that individuals with DS have stable, long-lasting perturbations in their cytokine levels similar to those in acute COVID-19. We conclude that DS can be considered to be a cytokinopathy.

Basal IL-6-mediated T cell activation

Cytometry by time-of-flight (CyTOF)-based immunophenotyping of whole blood from individuals with DS and age-matched HC individuals (Extended Data Fig. 2a–i) revealed that both CD4 and CD8 T cells in DS were skewed towards a memory phenotype (Fig. 2a,b and Extended Data Fig. 2d,e). There were fewer naive CD4 and CD8 T cells in individuals with DS, with a concurrent increase in CD4 central memory frequency[11,19]. We also uncovered baseline phosphorylated STAT3 (pSTAT3) in naive, activated and central memory CD4 T cells of individuals with DS, suggesting active cytokine signalling (Fig. 2c and Extended Data Fig. 3a).

STAT3 is a transcription factor that is activated downstream of multiple cytokines and growth factors[20]. When aberrantly phosphorylated, STAT3 contributes to lymphoproliferation, recurrent infections and increased autoimmunity including eczema, type 1 diabetes and hypothyroidism[21]. Clinically, Mendelian *STAT3* gain of function (GOF) substantially overlaps with DS, indicating that engagement of STAT3 in CD4 T cells in DS may contribute to immune pathogenesis. The molecular mechanism of disease in STAT3 GOF is still debated, but STAT3 attenuation of STAT5 phosphorylation is the leading hypothesis. In DS, basal phosphorylated STAT5 (pSTAT5) levels were unaffected in

CD4 T cells and even increased in CD4 TEMRAs (Extended Data Fig. 3b), suggesting that other factors are involved.

To test whether basal pSTAT3 is caused by cytokine signalling instead of an intrinsic STAT3 GOF in DS, we incubated whole blood with the FDA-approved Jak inhibitor tofacitinib. This treatment restored basal pSTAT3 to control levels, indicating that STAT3 activation in DS is Jak dependent (Fig. 2d and Extended Data Fig. 3c).

Given the detection of IFNα2 and IL-10 in a subset of individuals with DS (Fig. 1a), combined with the triplication of their cognate receptors (*IFNAR1*, *IFNAR2* and *IL10RB*) in DS, we hypothesized that basal pSTAT3 was due to heightened IFN-I or IL-10 response. However, IFN-I blockade did not affect basal pSTAT3 (Fig. 2e). Blocking IL-10 resulted in a modest decrease in pSTAT3 that was most pronounced in effector memory CD4 cells (Fig. 2f), suggesting that IL-10 may contribute to baseline signalling in DS. These blocking experiments were performed ex vivo, so the effect of these cytokines over time may not have been captured.

Having established that basal pSTAT3 is initiated upstream of Jaks and largely independent of IFN-I and IL-10, we turned to the IL-6–STAT3 axis. IL-6 is a potent inducer of pSTAT3 and was significantly elevated in DS (Fig. 1a,c). IL-6 blockade with the FDA-approved IL-6 receptor (IL-6R) inhibitor tocilizumab fully abrogated pSTAT3 (Fig. 2g and Extended Data Fig. 3d), indicating that IL-6 is a major mediator of baseline CD4 activation in DS. Furthermore, there was no hyper-response to exogenous IL-6 stimulation and no increase in IL-6R expression in DS (Extended Data Fig. 3e,f), confirming that elevated pSTAT3 is a result of high IL-6 rather than an intrinsic gain of expression or function in IL-6R. Together, these data implicate IL-6 signalling in ongoing CD4 T cell activation in DS and offer a more specific therapeutic target than Jak inhibition for consideration by physicians treating individuals with DS.

Notably, genetic alterations of STAT3 signalling also lead to a disturbance of B cell subsets, with a decrease in a recently described atypical B cell activation in adults with *STAT3* loss of function and a corresponding increase in these CD11c$^+$ B cells in *STAT3* GOF. Given our findings of basal pSTAT3 in DS T cells, the overlap in clinical manifestations between DS and STAT3 GOF, and previously documented B cell disturbances in DS[13], we further investigated the B cell compartment.

Increased CD11c$^+$Tbet^{high}CD21^{low} B cells

Total B cells were profoundly decreased in individuals with DS (Fig. 3a,b and Extended Data Fig. 4a) as seen previously[22,23]. All whole-blood samples were fixed before freezing, which enabled us to analyse all B cell populations including plasmablasts. The relative frequencies of B cell subsets were altered in DS: memory B cells were slightly decreased compared with HC individuals, whereas plasmablasts increased by almost threefold (Fig. 3a,c). Given the unexpected co-occurrence of elevated plasmablasts and depleted memory cells, combined with the simultaneous widespread elevation of pro-inflammatory cytokines and overactivation of CD4 T cells discussed above, we examined the possibility that B cells undergo inordinate activation in DS.

CD11c$^+$ B cells are thought to derive from naive B cells by cytokine and T cell stimulation and/or TLR engagement outside of germinal centres[24]. They are also characterized by high Tbet and low CD21 expression and can differentiate into plasmablasts[25]. The two major documented subsets of CD11c$^+$ B cells are CD27$^-$IgD$^+$ activated naive (aN) B cells and a subpopulation of class-switched CD27$^-$IgD$^-$ (double-negative (DN2)) B cells[26,27] (Extended Data Fig. 4b). This B cell activation is a hallmark of the heterogenous condition systemic lupus erythematosus (SLE). It has also been described in other autoimmune and autoinflammatory diseases including rheumatoid arthritis, ulcerative colitis and common variable immunodeficiency[25,27,28].

Our analysis revealed increased frequencies of CD11c$^+$ B cells in both the IgD$^+$ naive and double-negative (CD27$^-$IgD$^-$) compartments in individuals with DS compared with in HC individuals (Fig. 3d–f), at a frequency comparable to mild SLE (Fig. 3d–f). Elevation of previously described CD11c$^+$ intermediates (aN and DN2 B cells) was concurrent with a reduction in traditional B cell activation, evidenced by decreased frequencies in resting naive (rN) and DN1 B cells. The rN:aN and DN1:DN2 ratios were significantly lower in individuals with DS and SLE compared with in HC individuals (Fig. 3g). Despite the inherently variable cell frequencies in childhood, aN B cells were significantly higher in children with DS, and DN2 B cells trended upward (Extended Data Fig. 4c–g), demonstrating that this atypical B cell activation also occurs early in life.

We next performed high-dimensional analysis of functional markers of CD11c$^+$ B cells. As reported previously, both aN and DN2 B cells displayed upregulation of the IFNγ-induced transcription factor Tbet (Fig. 3h), and the surface receptors FAS (also known as CD95) and CD86 (also known as B7-2)[29] (Fig. 3i). These cells also exhibited lower expression of CD21 (Fig. 3j). Expression of chemokine receptors in CD11c$^+$ cells was consistent with previously described extrafollicular activation. The follicle-homing receptors CXCR5 and CCR7 were significantly downregulated in aN and DN2 B cells compared with in rN and DN1 B cells (Fig. 3k), concurrently with an increase in CXCR3 in aN cells (Fig. 3k), which suggests homing potential to inflamed tissues[30]. We also found elevated expression of a marker in aN B cells, CCR4 (Fig. 3k), a receptor that is classically associated with CD4 T_H2 cells and is thought to be involved in homing to the skin and lungs[31,32]. Further research is needed to elucidate whether CCR4 expression in these B cells can drive pathology.

Although CD11c$^+$ B cell frequency in DS was not correlated with age[27,33] (Extended Data Fig. 4h), it was correlated with total cytokine levels (Extended Data Fig. 4i), specifically IL-6 levels (Fig. 3l), as well with the frequency of circulating T follicular helper 1/17 (cT_{FH}1/17) cells (Fig. 3m), as in other primary immunodeficiencies[28]. This suggests that cytokines and activated CD4 T cells contribute to putative extrafollicular B cell differentiation[24]. Finally, CD11c$^+$ B cell frequency was correlated with the number of autoimmune manifestations in our cohort (Extended Data Fig. 4j), indicating their potential role in autoimmunity.

In conclusion, CD11c$^+$ B cells are a prominent component of the abnormal B cell response in DS. On the basis of our research and previous work in other inflammatory conditions[25,28], we hypothesize that this dysregulated B cell response weakens immune tolerance and, ultimately, contributes to autoimmunity in DS.

Ex vivo naive B cell differentiation

CD11c$^+$ B cells can differentiate into antibody-secreting plasmablasts with little to no affinity maturation, potentially explaining their association with autoimmunity[34]. In vitro, we confirmed that stimulated CD11c$^+$ B cells sorted from HC peripheral blood mononuclear cells (PBMCs) differentiated into plasmablasts and secreted immunoglobulin G (IgG), in contrast to treated naive cells (Fig. 4a,b). This was potentiated by the addition of IFNγ (Fig. 4a), as previously published[27]. In whole blood, CD11c$^+$ B cell frequency was positively correlated with that of plasmablasts (Fig. 4c). We also confirmed that total IgG is elevated in DS plasma compared to HC plasma[13] (Extended Data Fig. 4k).

We next performed experiments to test the mechanisms underlying increased B cell activation in DS. To first address whether the cytokines in DS plasma could induce this differentiation, we co-incubated naive or total B cells sorted from control PBMCs in extrafollicular-stimulating conditions[27] in the presence of IgG-depleted plasma derived from HC individuals or individuals with DS. There was significantly more plasmablast differentiation in the presence of DS plasma compared to HC plasma (Fig. 4d and Extended Data Fig. 5a). Exogenous cytokine treatment could also influence plasmablast differentiation: IFNα and IFNγ potentiated differentiation, whereas IL-4 restricted it[35,36] (Extended Data Fig. 5b–d). Individual blocking of IL-6, IFN-I, IFNγ and TNF in donor plasma had no effect, whereas Jak inhibitors inhibited plasmablast differentiation, suggesting that a few cytokines are involved (Extended Data Fig. 5f). Blocking a combination of four cytokines (IL-6, IFN-I, IFNγ and TNF) reversed the DS plasma-induced augmentation of plasmablast

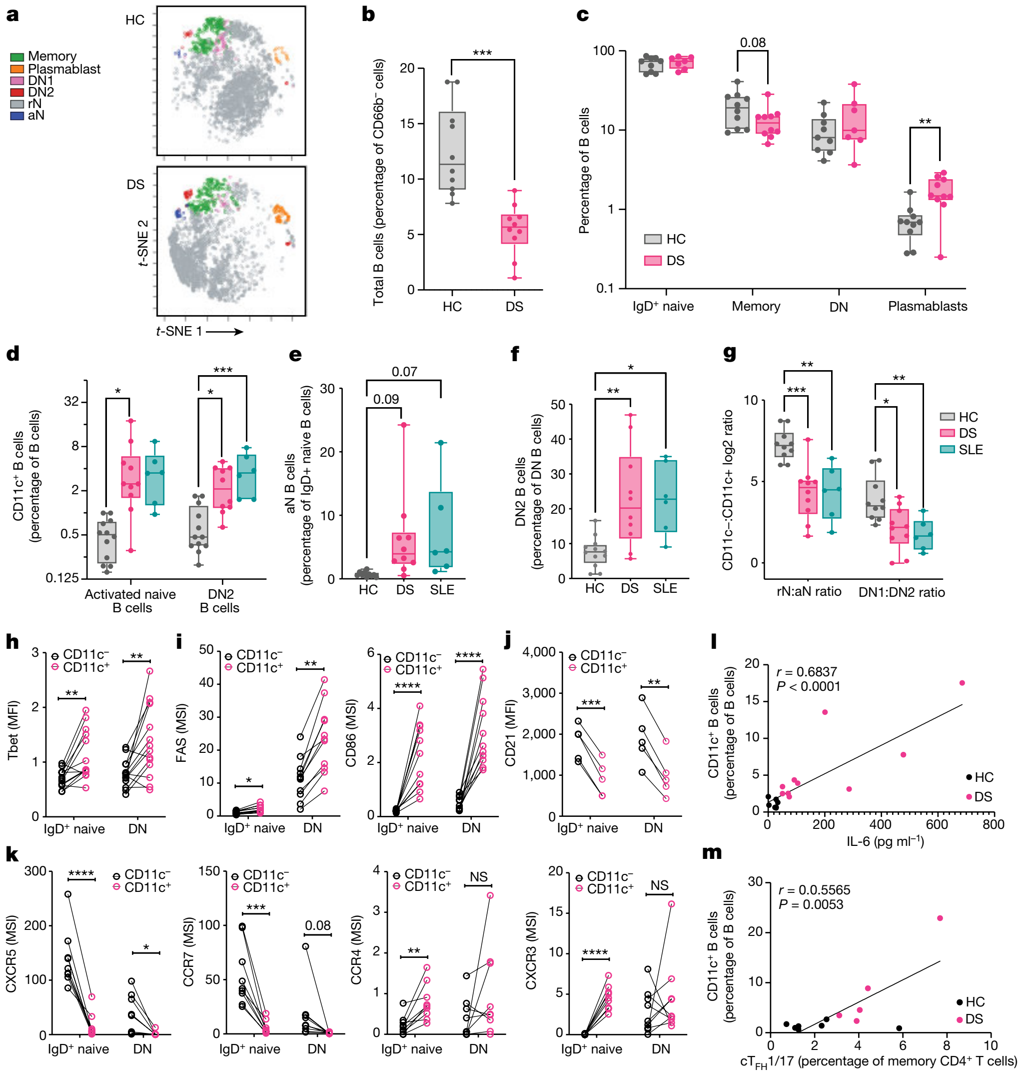


Fig. 3 | Increased frequency of CD11c$^+$Tbet^{high}CD21^{low} B cells in DS. **a**, Representative *t*-distributed stochastic neighbour embedding (*t*-SNE) analysis of a fixed number of B cells to illustrate subset distribution in whole blood from adults with DS and age-matched HC individuals. **b**,**c**, The frequency in adults with DS ($n = 10$) and age-matched HC individuals ($n = 10$) of total B cells expressed as the percentage of CD66b$^-$ cells (non-granulocytes) (**b**) and B cell subsets expressed as the percentage of total B cells (**c**). **d**–**g**, The frequency in HC adults ($n = 10$), adults with DS ($n = 10$) and patients with SLE ($n = 6$) of CD11c$^+$ B cells in the IgD$^+$ naive (CD27$^-$CD38^{low}IgD$^+$) or DN (CD27$^-$CD38^{low}IgD$^-$) compartments expressed as the percentage of total (**d**), IgD$^+$ naive (**e**) or DN (**f**) B cells, and the log$_2$-transformed ratios of CD11c$^+$ subsets to CD11c$^-$ subset (rN:aN and DN2:DN1) (**g**). **h**–**k**, Intracellular Tbet expression (**h**), and surface expression of FAS and CD86 (**i**), CD21 (**j**) and CXCR5, CCR7, CCR4 and CXCR3 (**k**) in naive and DN B cells from both HC individuals ($n = 2$–4) and individuals with DS ($n = 3$–10). MSI, mean signal intensity; MFI, mean fluorescence intensity. **l**,**m**, Correlation of CD11c$^+$ B cells and circulating IL-6 (**l**) and cT_{FH}1/17 (**m**). *r*, Pearson correlation coefficient. In **b**–**g**, the whiskers denote the minimum and maximum values, the box limits denote quartile 1 to quartile 3, and the centre bar denotes the mean. Significance was assessed using two-tailed unpaired *t*-tests (**b** and **h**–**k**) and one-way ANOVA with Tukey's post hoc analysis (**d**–**g**).

differentiation (Fig. 4e). Thus, the cytokines present in DS plasma—at least, IL-6, IFN-I, IFNγ and TNF—are drivers of naive B cell differentiation into plasmablasts.

Given the mounting evidence in vivo that T cells drive extrafollicular B cell activation[24,28], together with our findings of basal IL-6 signalling in DS CD4 T cells, we tested whether T cells contribute to this B cell response. Previous studies have demonstrated that T_H1-polarized T cells can induce extrafollicular differentiation of naive B cells in vitro[33,37]. When we modelled DS CD4 T cell activation with exogenous IL-6, co-culture of naive B cells and T cells pretreated with IL-2 and IL-6 resulted in plasmablast differentiation equivalent to that of T_H1–B cell co-cultures (Fig. 4f). Exogenous IL-6 alone did not affect plasmablast differentiation (Extended Data Fig. 5b). Furthermore, T_H1-cell- and IL-6-primed CD4 T cells induced CD11c expression in co-cultured B cells (Fig. 4g), together with CD21 downregulation (Fig. 4h) and a slight increase in Tbet (Extended Data Fig. 5g). Although IFNγ is thought to have a major role in the T_H1-mediated atypical B cell differentiation and was indeed present in the T_H1 supernatants, we did not detect IFNγ in the IL-6 conditions (Extended Data Fig. 5h), which indicates that other cytokines drive B cell activation in these conditions. Together, these results demonstrate that cytokines and T cells in combination can drive an extrafollicular B cell response.

To better replicate the physiological conditions of DS, we performed these experiments with pre-incubation of T cells in plasma derived from individuals with DS. We found that these cells induced increased plasmablast differentiation of naive B cells compared with T cells incubated in control plasma (Extended Data Fig. 5i). Finally, to determine whether activated CD4 T cells of individuals with DS are poised to induce differentiation of naive B cells, we isolated CD4 T cells from control individuals ($n = 4$) and individuals with DS ($n = 2$) and co-cultured them with CD11c$^-$ naive B cells in syngeneic co-cultures. Without exogenous polarization, T cells from individuals with DS induced more CD11c

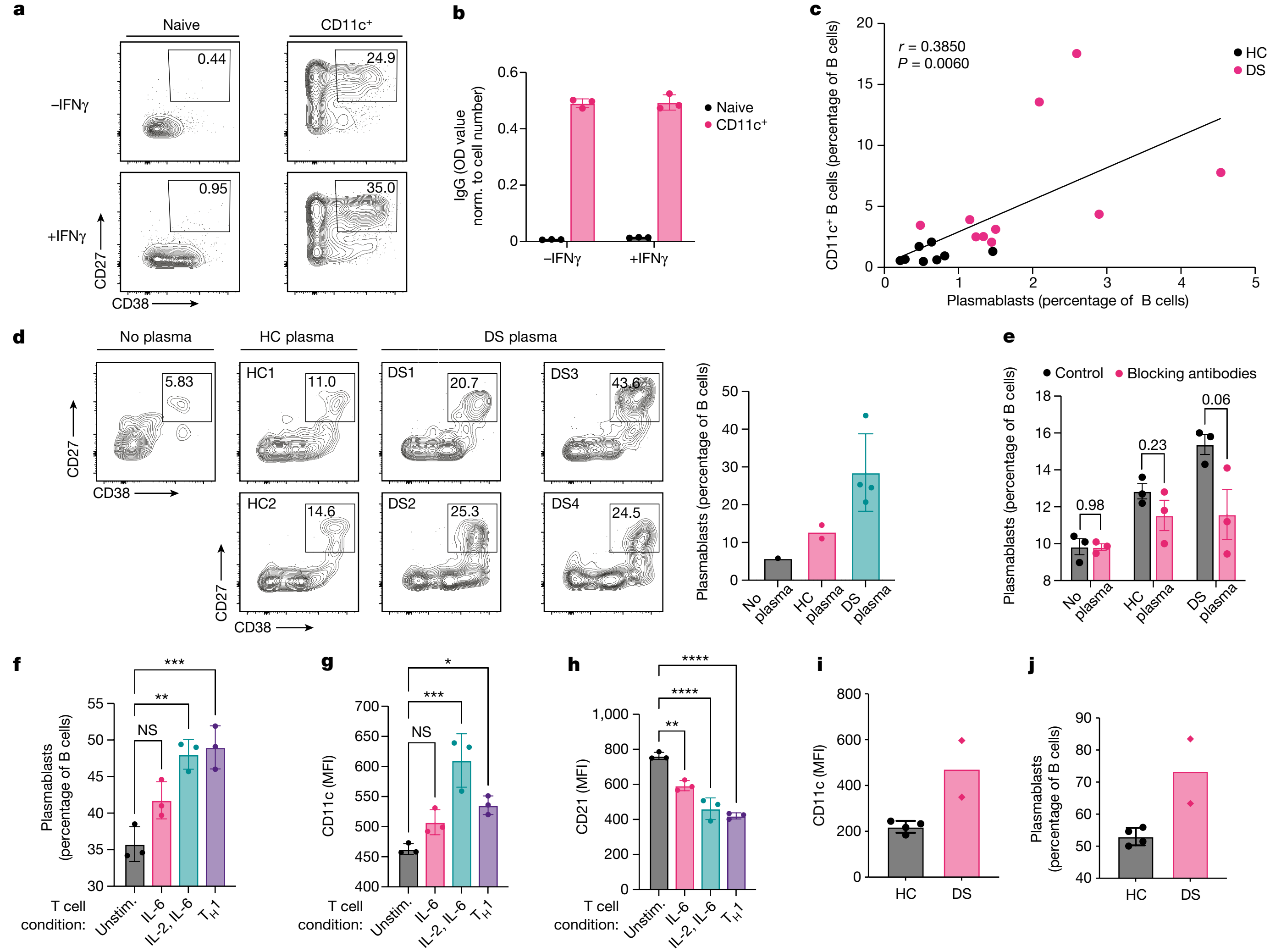


Fig. 4 | B cell activation by DS plasma and stimulated T cells. a,**b**, Plasma cell differentiation (**a**) and secreted IgG in the supernatant (**b**) after culture for 4 days of sorted HC naive or CD11c^{+} B cells in the presence of BAFF, IL-2, IL-10, IL-21, the TLR7/8 ligand R848 with or without IFNγ. $n = 3$ biologically independent samples. Norm., normalized. **c**, Correlation of CD11c^{+} B cell and plasmablast frequencies in adults DS ($n = 12$) and age-matched HC individuals ($n = 8$). **d**, Plasma cell differentiation after culture for 6 days of sorted HC naive B cells in the presence of BAFF, IL-2, IL-10, IL-21, R848 and IgG-depleted plasma from HC individuals ($n = 2$) or individuals with DS ($n = 4$). **e**, Magnetic-activated cell sorting (MACS)-isolated naive B cells from a healthy donor were cultured for 3 days with BAFF, IL-2, IL-21, R848, anti-IgM and plasma of HC individuals or individuals with DS in the presence of a combination of antibodies blocking IFN-I, IFN-II, IL-6 and TFNα signalling, in triplicates. Significance was assessed using two-tailed unpaired t-tests. Data are mean ± s.e.m. The results are representative of two independent experiments with $n = 2$ donors per group. **f**–**h**, Co-cultures containing T cells activated with IL-6, IL-2, both or polarized into T_H1 cells with IL-2, IL-12 and anti-IL-4 together with MACS-isolated naive B cells from the same donor, run in triplicates. The frequency of plasmablasts (**f**) and extracellular CD11c induction (**g**) and downregulation of CD21 expression (**h**) in non-plasmablast B cells after co-culture for 3–6 days. Significance was assessed using one-way ANOVA with Tukey's post hoc analysis. **i**,**j**, CD11c induction (**i**) and plasmablast differentiation (**j**) in naive CD11c^{-} B cells isolated from controls ($n = 4$) or individuals with DS ($n = 2$) after 3 days of co-culture with CD4 T cells isolated from the same donors (syngeneic cultures). For **b**, **d** and **f**–**j**, data are mean ± s.d.

expression and plasmablast differentiation than those of HC individuals (Fig. 4i,j). These data indicate that cytokine milieu and steady-state cellular activation contribute to atypical B cell activation in DS.

Autoimmune features in CD11c^{+} B cells

Next, we looked at immunoglobulin isotype expression to further ascertain the antibody-secreting potential and the naivety of CD11c^{+} B cells. The frequency of IgD^{+}CD11c^{+} B cells was intermediate between naive and memory B cells (Fig. 5a). The proportion of IgA^{+}CD11c^{+} B cells was similar to memory B cells (Fig. 5b), demonstrating that a portion of these atypical cells have gone through class switching and are therefore probably antigen experienced. We did not detect significant differences in isotype usage between HC and DS CD11c^{+} B cells (Extended Data Fig. 6a).

To assess the clonality of CD11c^{+} B cells, we performed B cell receptor (BCR) sequencing (BCR-seq) of DNA from sorted naive, CD11c^{+} and memory B cells from individuals with DS ($n = 6$) and age-matched control individuals ($n = 6$) at steady state (Extended Data Fig. 6b,c). In naive cells, around 99% of BCRs were unique, whereas, in CD11c^{+} cells and memory cells, up to 7% of BCRs were expanded (Fig. 5c), suggesting that, like memory B cells, CD11c^{+} B cells undergo clonal expansion (or are the result of clonal expansion) rather than non-specific stimulation. There was no difference in clonality in CD11c^{+} B cells between HC individuals and individuals with DS, as evidenced by similar fractions of non-unique BCR templates (Fig. 5c) and similar Simpson diversity index metrics (Extended Data Fig. 6d). The complementarity determining region 3 (CDR3) repertoires of CD11c^{+} and memory B cell subsets overlapped at the nucleotide and amino acid levels (Extended

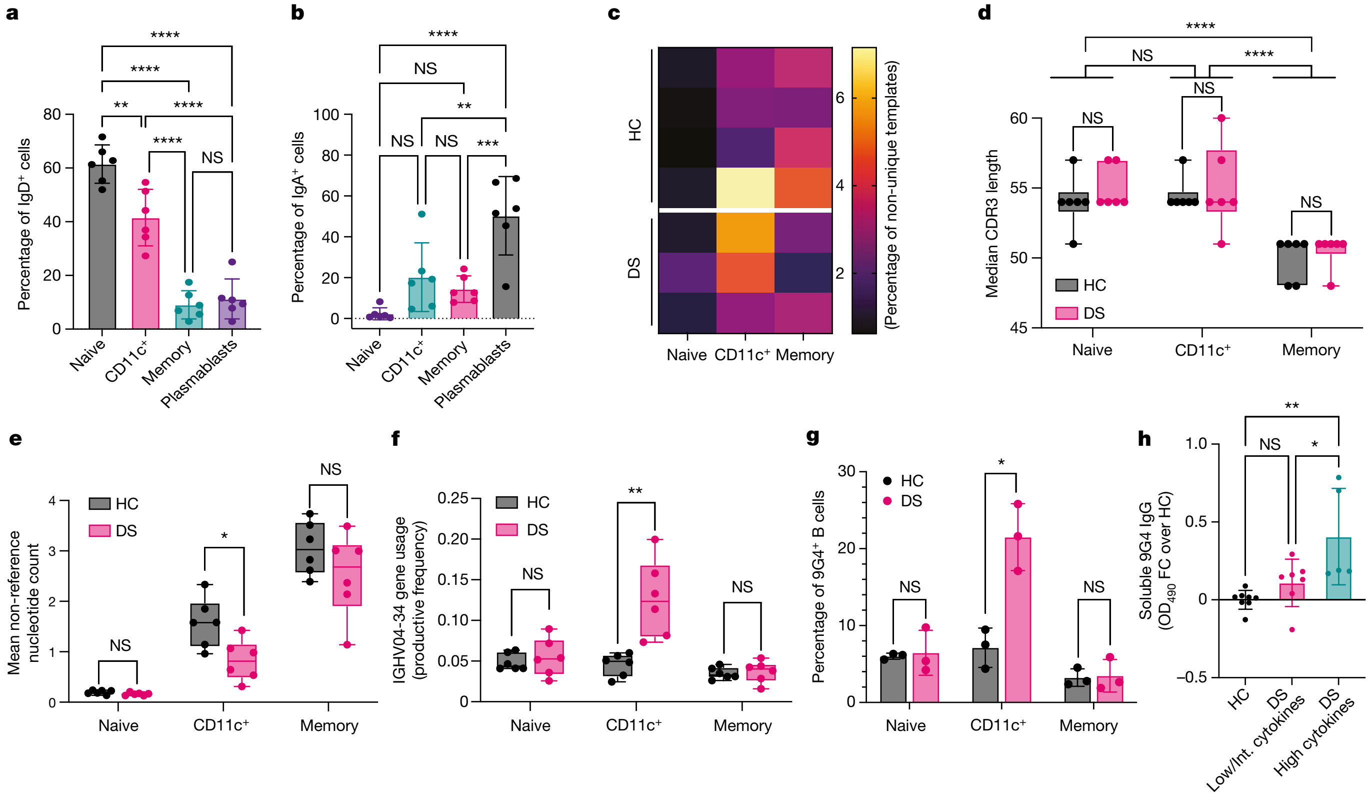


Fig. 5 | CD11c⁺ B cells from individuals with DS are more prone to autoreactivity. a,**b**, Expression of IgD (**a**) and IgA (**b**) in naive, CD11c⁺ and memory B cells and plasmablasts from adults with DS and age-matched HC individuals. $n = 3$ each. **c**–**h**, BCR sequencing analysis of genomic DNA isolated from sorted naive, CD11c⁺ and memory B cells from controls ($n = 6$) and individuals with DS ($n = 6$). **c**, The fraction of productive BCRs represented more than once in each sample. Individuals from whom a sample had fewer than 1,000 productive templates were excluded. **d**, CDR3 length of productive BCRs in B cells subsets in HC and DS, expressed as the number of nucleotides (nt). **e**, The mean number of nucleotides different from reference in the V gene of productive BCRs in B cells subsets in HC and DS. **f**, The frequency of productive BCRs that were aligned to the *IGHV4-34* gene in each sample. **g**, 9G4 surface expression in naive, CD11c⁺ and memory B cells from HC individuals ($n = 3$) and individuals with DS ($n = 3$). **h**, ELISA quantification of 9G4 antibodies in the plasma of HC individuals ($n = 8$), and individuals with DS in the low/medium ($n = 7$) and high ($n = 5$) cytokine groups, expressed as the fold change over HC individuals. OD_{490}, optical density at 490 nm. For **a**, **b**, **g** and **h**, data are mean ± s.d. For **d**–**f**, the whiskers denote the minimum and maximum values, the box limits denote quartile 1 to quartile 3, and the centre bar denotes the mean. Significance in **a** and **h** and significance between cell subsets in **d** was assessed using one-way ANOVA with Tukey's post-hoc analysis. Significance in **d**–**g** between the HC and DS groups was assessed using two-tailed paired *t*-tests.

Data Fig. 6e,f), demonstrating that cells in these two phenotypically defined subsets derive from the same lineage. In future studies, BCR-seq analysis of donors immunized with a known antigen will enable closer examination of clonal expansions of these B cells.

The median CDR3 length of CD11c⁺ B cells was similar to that of naive cells and significantly longer than that of memory cells (Fig. 5d). It was not significantly different between these cells in individuals with DS and HC individuals. Increased CDR3 length in antibody-secreting cells is associated with antibody polyreactivity and autoimmunity[38], suggesting that these CD11c⁺ B cells are prone to autoreactivity.

The number of non-reference nucleotides in the V genes of CD11c⁺ B cells was intermediate between that of naive and memory B cells (Fig. 5e). Notably, CD11c⁺ B cells in individuals with DS had significantly fewer non-reference nucleotides compared with in HC individuals. This may indicate a comparative lack of somatic hypermutation, which could lead to a broader, more non-specific humoral response. Our method was limited to CDR3 sequencing; thus, characterization of the full antigen-binding region is still needed to ascertain the true somatic hypermutation rate.

Analysis of BCR V gene usage (Extended Data Fig. 6g) revealed that CD11c⁺ B cell expansion is probably more autoreactive in individuals with DS than in HC individuals. There was a significantly higher usage of the *IGHV4-34* gene, which is associated with autoreactivity[27,38], in CD11c⁺ B cells in DS (Fig. 5f). We confirmed these BCR-seq results using flow cytometry: 9G4 idiotype antibodies encoded by this *IGHV4-34* gene segment were more highly expressed in these atypical B cells in DS (Fig. 5g). In vitro stimulation of sorted CD11c⁺ B cells into antibody-secreting cells led to higher 9G4 secretion than from memory B cells (Extended Data Fig. 4h), further accentuating the link between these rare B cells and autoimmune potential. Circulating 9G4 antibodies were also elevated in the plasma of individuals with DS and were more abundant in the high-cytokine individuals with DS than in the low/intermediate groups (Fig. 5h). 9G4 antibodies are known contributors of autoimmunity in SLE, displaying specificity for nuclear antigens, dsDNA and apoptotic cells[39,40]. In conclusion, CD11c⁺ B cells in DS are present at a higher frequency and are more likely to exhibit features of self-reactive BCRs.

Autoantibodies are enriched in DS

Given clinical autoimmunity in DS and our findings of cytokine dysregulation and autoimmune-prone B cells, we hypothesized that trisomy 21 results in the generation of autoantibodies. To characterize the DS autoreactive repertoire, we assessed the plasma IgG and IgA reactivity of individuals with DS ($n = 5$), age-matched healthy individuals ($n = 4$), 3 individuals with immunodysregulation polyendocrinopathy enteropathy X-linked syndrome (IPEX) and 1 individual with autoimmune

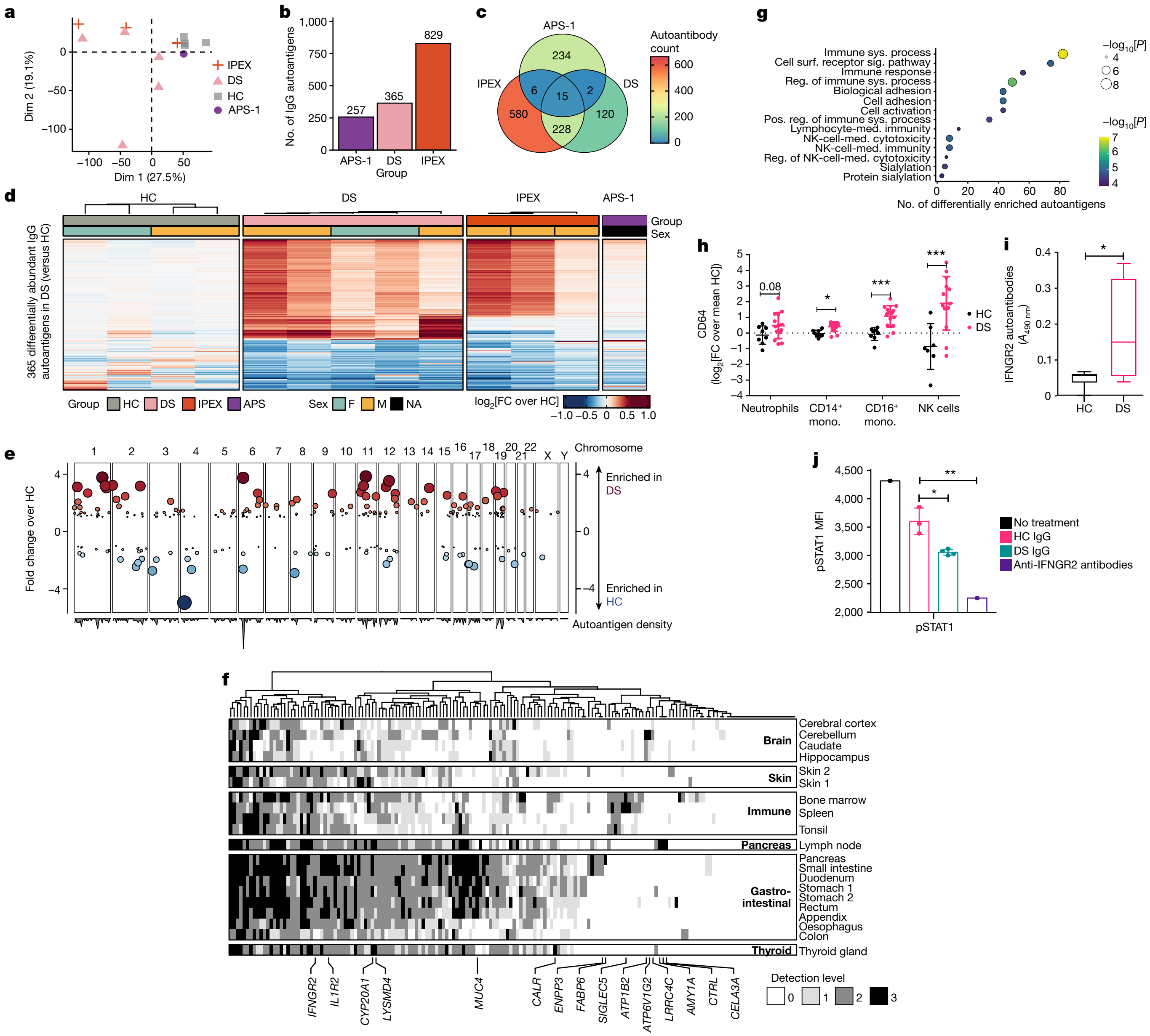


Fig. 6 | Autoantibody repertoire in individuals with DS. a, PCA analysis of the HuProt IgG dataset for adults with DS ($n = 5$), age-matched HC individuals ($n = 4$), and patients with IPEX ($n = 3$) and APS-1 ($n = 1$). **b**, The number of IgG autoantigens enriched at least twofold in individuals with DS, IPEX and APS-1 compared with HC individuals. **c**, Enriched IgG autoantigens overlapping between disease groups. **d**, Enriched IgG autoantigens in HC individuals and in individuals with DS, IPEX and APS-1. The colour intensity corresponds to the log_2-transformed fold change expression value relative to the mean of healthy adult controls. F, female; M, male; NA, unknown. **e**, Chromosomal expression pattern of IgG autoantigens enriched in DS. **f**, Gene expression pattern of IgG autoantigens enriched in DS ($n = 5$) according to the Human Protein Atlas. **g**, GO analysis of IgG autoantigens enriched in DS ranked by the number of autoantigens found to be enriched in the associated gene set. The dot size and colour intensity correspond to the FDR-adjusted *P* value. Med., mediated; NK, natural killer; reg., regulation; sig, signalling; surf, surface; sys., system. **h**, The surface expression of CD64 in monocytes, natural killer cells from individuals with DS ($n = 14$) and age-matched HC individuals ($n = 8$). Mono, monocytes. **i**, ELISA analysis of anti-IFNGR2 autoantibodies in the plasma from adults with DS ($n = 4$) and age-matched HC individuals ($n = 7$). **j**, Neutralizing IFNγ signalling in THP-1 cells by IgG fraction of plasma from adults with DS ($n = 3$), age-matched HC individuals ($n = 3$) or recombinant anti-IFNGR2 antibody. Significance was assessed using one-way ANOVA with Tukey's post hoc analysis. For **h** and **i**, significance was assessed using two-tailed unpaired *t*-tests. For **h** and **j**, data are mean ± s.d. For **i**, the whiskers denote the minimum and maximum values, the box limits denote quartile 1 to quartile 3, and the centre bar denotes the mean.

polyglandular syndrome type 1 (APS-1) (disease controls; $n = 4$) against >21,000 conformationally intact human proteins (CDI HuProt protein microarray) (Fig. 6a and Extended Data Fig. 7a–d).

Principal component analysis (PCA) grouped DS samples together and away from HC individuals (Fig. 6a), indicating differential self-antigen binding. The DS samples clustered with samples from patients with IPEX ($n = 3$), and away from the sample from the individual with APS-1 ($n = 1$) (caused by mutation in *AIRE*), although the low sample number is an important caveat (Fig. 6a).

Analysis of differentially abundant autoantigens in individuals with DS compared with HC individuals (log_2-transformed fold change > 1, $P < 0.05$) yielded 365 proteins, in contrast to 257 and 829 proteins for APS-1 and IPEX, respectively (Fig. 6b). This indicates that autoimmunity in trisomy 21 is on par with bona fide autoimmune diseases.

We observed little overlap of autoantigens between DS and APS-1 (2 out of 257), whereas a large proportion of the DS autoantigens overlapped with those in IPEX (228 out of 829) (Fig. 6c,d), a syndrome of severe multiorgan inflammation caused by defects in T regulatory (T_{reg}) cells. The highly similar autoantibody repertoires points to T cell overactivation as a possible driver of autoreactive B cells in DS. Consistent with this finding, absolute T_{reg} cell counts were decreased in individuals with DS compared with in HC individuals, despite similar T_{reg} cell frequency[41] (Extended Data Fig. 2f,g). Finally, 120 autoantigens were unique to DS, including metabolic (TPH1, ATP5F1, PLPP2), cell signalling (RRAGC, GPR143, PTGER4), immune signalling (DTX4, CCL11, TROVE2) and neuronal pathway intermediates (ABAT, ATP6V1G2, KCTD7) (Fig. 6c,d and Supplementary Data 1 and 2).

The autoantigens enriched in DS were encoded across all 23 chromosomes (Fig. 6e). They included proteins expressed at sites of clinical autoimmunity in DS, such as the gastrointestinal tract (MUC4, ENPP3, FABP6) and the pancreas (AMY1A, CELA3A, CTRL) (Fig. 6f). The four individuals with DS with hypothyroidism assayed did not have known anti-thyroid autoantibodies in contrast to previously described individuals with DS[42] but, instead, had antibodies directed against proteins non-specifically expressed in the thyroid (CALR, LYSMD4, CYP20A1) (Fig. 6f). We also saw an enrichment of autoantigens expressed in the central nervous system (CNS) (ATP1B2, LRRC4C, ATP6V1G2) (Fig. 6f). Although antibodies cannot cross the intact blood–brain barrier, pathology can be mediated through a weakened blood–brain barrier[43]. Erroneous expression of normally neuronally restricted markers through altered gene expression in DS is another potential mechanism[44]. Given the prominent central nervous system features of DS, further investigation is warranted. Finally, immune autoantigens were prominently enriched in DS, including IFNGR2, TLR9, CD1C and IL-1R2 (Fig. 6f,g and Extended Data Fig. 7e). Immune system process, cell-surface receptor signalling and immune response were the most significantly enriched Gene Ontology (GO) terms (Fig. 6g). Abundant autoantibodies directed against the immune system may promote further immune dysregulation.

On the cellular side, we detected increased expression of high-affinity FCGR1A (also known as CD64) in monocytes and natural killer cells in DS (Fig. 6h). This receptor can engage bound antibodies and immune complexes to trigger potent inflammation and tissue injury[45]. Cells in individuals with DS may be poised to cause tissue damage based on the observed expression patterns.

We next validated the HuProt array results using an enzyme-linked immunosorbent assay (ELISA), confirming the presence of anti-IFNGR2, MSTN and ATP6V1G2 autoantibodies in individuals with DS but not in HC individuals (Fig. 6i and Extended Data Fig. 7f). We then functionally tested anti-IFNGR2 autoantibodies (enriched 2.7-fold in DS, false-discovery rate (FDR) = 0.01): compared to HC IgG, DS IgG resulted in a significant decrease in the response to IFNγ as measured by pSTAT1 (Fig. 6j). This demonstrates that immune-targeting autoantibodies in individuals with DS can directly inhibit the activity of cytokines. As genetic defects in IFNγ signalling cause susceptibility to mycobacterial disease[46], anti-IFNGR2 autoantibodies may predispose individuals with DS to mycobacterial disease, which should be investigated in DS especially as individuals age.

Given the increased susceptibility to severe COVID-19 in individuals with DS[47], we examined anti-IFN-I autoantibodies. These were modestly enriched in individuals with DS compared with HC individuals in the CDI array, but much less so than in APS-1 (Extended Data Fig. 7g). We confirmed anti-IFNα2 and anti-IFNω autoantibody enrichment in DS by Gyros assay, at titres over tenfold lower than that of APS-1 samples (Extended Data Fig. 7h). Plasma from individuals with DS with or without COVID-19 was insufficient to inhibit IFN-I, whereas samples from the individual with APS-1 fully abrogated IFN-I signalling. We noted only mild neutralization of very low amounts of IFNα2 (100 pg ml^{-1}) by DS plasma (Extended Data Fig. 7i).

In conclusion, we detected a prominent autoantibody landscape in the plasma of individuals with DS. Together with our findings of B and T cell activation, we hypothesize that this broad humoral response predisposes individuals with DS to form high-affinity, tissue-specific antibodies that manifest clinically as autoimmune disease.

Discussion

Autoantibodies in DS have long been implicated in many of the syndrome's features, such as thyroid disease, type 1 diabetes and even cognitive decline[42,48–50]. Here we delineate the predisposition to autoreactivity and identify the autoantibody repertoire in DS (Extended Data Fig. 8). Two of the dysregulated immune features that we detected—the cytokine landscape and $CD11c^+$ B cell frequency—had a direct moderate association with clinical immune manifestations ($P = 0.0141$ by ANOVA and $P = 0.0795$ by r^2 test, respectively). Perhaps dampening chronic steady-state inflammation in DS could prevent the accumulation of self-reactive autoantibodies and lessen disease burden.

We identified an increased frequency of $CD11c^+Tbet^+CD21^{low}$ B cells as a hallmark of immune dysregulation in DS and established a link between these cells and the elevated plasmablast frequency and increased IgG in DS plasma. Others have also implicated these cells in the production of self-reactive antibodies[34], which may partly explain autoimmune susceptibility in DS. Beyond the increased frequency of these intrinsically pro-inflammatory cells in DS, BCR sequencing revealed that $CD11c^+$ cells are more likely to have self-reactive features. Thus, in DS, these putatively extrafollicular B cells are skewed towards autoimmunity in both quantity and quality. Their direct involvement in the generation of autoantibodies and their potential damage at the tissue level merit further investigation.

We uncovered activation of naive B cells not only by cytokines in DS plasma but also by CD4 T cells with basal pSTAT3. Patients with *STAT3* loss-of-function mutations have fewer $CD11c^+$ B cells, whereas those with *STAT3* GOF mutations have more[28], like individuals with DS. Although it is unclear whether the exaggerated naive B cell activation in *STAT3* GOF is due to intrinsic STAT3-mediated signalling in B cells, stimulation by overactive T cells or both, we hypothesize that baseline T cell activation in DS has an important role in this B cell response. This notion is further emphasized by the similarity in autoantibody repertoires between DS and IPEX, an autoimmune disease mediated by lack of T cell regulation. Our in vitro experiments demonstrate that IL-6 activation of T cells is sufficient to drive the differentiation of naive B cells into $CD11c^+$ B cells and plasmablasts. Although the physiological implications of this model must be firmly tested, these data provide a mechanistic link between IL-6 activation of T cells and this atypical B cell response. Thus, resolving basal T cell activation in DS with a Jak inhibitor or IL-6 blockade is a promising therapeutic avenue to temper hyperinflammation and the autoimmune feed-forward loop that we describe.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-05736-y.

1. Bull, M. J. Down syndrome. *N. Engl. J. Med.* **382**, 2344–2352 (2020).
2. Ryan, C., Vellody, K., Belazarian, L. & Rork, J. F. Dermatologic conditions in Down syndrome. *Pediatr. Dermatol.* **38**, 14731 (2021).
3. Down, J. L. H. Observations on an ethnic classification of idiots. *Lond. Hosp. Rep.* **3**, 259–262 (1866).
4. Mai, C. T. et al. National population-based estimates for major birth defects, 2010–2014. *Birth Defects Res.* **111**, 1420–1435 (2019).

5. Antonarakis, S. E. et al. Down syndrome. *Nat. Rev. Dis. Prim.* **6**, 9 (2020).
6. Rivelli, A. et al. Prevalence of endocrine disorders among 6078 individuals with Down syndrome in the United States. *J. Patient Cent. Res. Rev.* **9**, 70–74 (2022).
7. Kong, X.-F. et al. Three copies of four interferon receptor genes underlie a mild type I interferonopathy in Down syndrome. *J. Clin. Immunol.* **40**, 807–819 (2020).
8. Sullivan, K. D. et al. Trisomy 21 consistently activates the interferon response. *eLife* **5**, e16220 (2016).
9. Malle, L. et al. Excessive negative regulation of type I interferon disrupts viral control in individuals with Down syndrome. *Immunity* **55**, 2074–2084 (2022).
10. Giménez-Barcons, M. et al. Autoimmune predisposition in Down syndrome may result from a partial central tolerance failure due to insufficient intrathymic expression of AIRE and peripheral antigens. *J. Immunol.* **193**, 3872–3879 (2014).
11. Araya, P. et al. Trisomy 21 dysregulates T cell lineages toward an autoimmunity-prone state associated with interferon hyperactivity. *Proc. Natl Acad. Sci. USA* **116**, 24231–24241 (2019).
12. Cossarizza, A. et al. Age-related expansion of functionally inefficient cells with markers of natural killer activity in Down's syndrome. *Blood* **77**, 1263–1270 (1991).
13. Verstegen, R. H. J. & Kusters, M. A. A. Inborn errors of adaptive immunity in Down syndrome. *J. Clin. Immunol.* **40**, 791–806 (2020).
14. Taft, J. & Bogunovic, D. The Goldilocks zone of type I IFNs: lessons from human genetics. *J. Immunol.* **201**, 3479–3485 (2018).
15. Rea, I. M. et al. Age and age-related diseases: role of inflammation triggers and cytokines. *Front. Immunol.* **9**, 586 (2018).
16. Zhang, Y. et al. Aberrations in circulating inflammatory cytokine levels in patients with Down syndrome: a meta-analysis. *Oncotarget* **8**, 84489–84496 (2017).
17. Huggard, D. et al. Increased systemic inflammation in children with Down syndrome. *Cytokine* **127**, 154938 (2020).
18. Junttila, I. S. Tuning the cytokine responses: an update on interleukin (IL)-4 and IL-13 receptor complexes. *Front. Immunol.* **9**, 888 (2018).
19. Lambert, K. et al. Deep immune phenotyping reveals similarities between aging, Down syndrome, and autoimmunity. *Sci. Transl. Med.* **14**, 4888 (2022).
20. Milner, J. D. et al. Early-onset lymphoproliferation and autoimmunity caused by germline STAT3 gain-of-function mutations key points. *Blood* **21**, 25 (2015).
21. Fabre, A. et al. Clinical aspects of STAT3 gain-of-function germline mutations: a systematic review. *J. Allergy Clin. Immunol. Pract.* **7**, 1958–1969 (2019).
22. Verstegen, R. H. J. et al. Defective B-cell memory in patients with Down syndrome. *J. Allergy Clin. Immunol.* **134**, 1346–1353 (2014).
23. Dieudonné, Y. et al. Immune Defect in adults with down syndrome: insights into a complex issue. *Front. Immunol.* **11**, 840 (2020).
24. Song, W. et al. Development of Tbet- and CD11c-expressing B cells in a viral infection requires T follicular helper cells outside of germinal centers. *Immunity* **55**, 290–307 (2022).
25. Karnell, J. L. et al. Role of CD11c^{+} T-bet^{+} B cells in human health and disease. *Cell. Immunol.* **321**, 40–45 (2017).
26. Woodruff, M. C. et al. Extrafollicular B cell responses correlate with neutralizing antibodies and morbidity in COVID-19. *Nat. Immunol.* **21**, 1506–1516 (2020).
27. Jenks, S. A. et al. Distinct effector B cells induced by unregulated toll-like receptor 7 contribute to pathogenic responses in systemic lupus erythematosus. *Immunity* **49**, 725–739 (2018).
28. Keller, B. et al. The expansion of human T-bet^{high}CD21^{low} B cells is T cell dependent. *Sci. Immunol.* **6**, 52 (2021).
29. Li, H., Borrego, F., Nagata, S. & Tolnay, M. Fc receptor-like 5 expression distinguishes two distinct subsets of human circulating tissue-like memory B cells. *J. Immunol.* **196**, 4064–4074 (2016).
30. Sokol, C. L. & Luster, A. D. The chemokine system in innate immunity. *Cold Spring Harb. Perspect. Biol.* **7**, a016303 (2015).
31. Vestergaard, C. et al. A Th2 chemokine, TARC, produced by keratinocytes may recruit CLA^{+}CCR4^{+} lymphocytes into lesional atopic dermatitis skin. *J. Invest. Dermatol.* **115**, 640–646 (2000).
32. Mikhak, Z., Strassner, J. P. & Luster, A. D. Lung dendritic cells imprint T cell lung homing and promote lung immunity through the chemokine receptor CCR4. *J. Exp. Med.* **210**, 1855 (2013).
33. Zumaquero, E. et al. IFNγ induces epigenetic programming of human T-bet^{hi} B cells and promotes TLR7/8 and IL-21 induced differentiation. *eLife* **8**, e41641 (2019).
34. Wang, S. et al. IL-21 drives expansion and plasma cell differentiation of autoreactive CD11c^{hi} T-bet + B cells in SLE. *Nat. Commun.* **9**, 1758 (2018).
35. Jego, G. et al. Plasmacytoid dendritic cells induce plasma cell differentiation through type I interferon and interleukin 6. *Immunity* **19**, 225–234 (2003).
36. Naradikian, M. S. et al. Cutting edge: IL-4, IL-21, and IFN-γ interact to govern T-bet and CD11c expression in TLR-activated B cells. *J. Immunol.* **197**, 1023–1028 (2016).
37. Stone, S. L. et al. T-bet transcription factor promotes antibody-secreting cell differentiation by limiting the inflammatory effects of IFN-γ on B cells. *Immunity* **50**, 1172–1187 (2019).
38. Bashford-Rogers, R. J. M. et al. Analysis of the B cell receptor repertoire in six immune-mediated diseases. *Nature* **574**, 122–126 (2019).
39. Tipton, C. M., Hom, J. R., Fucile, C. F., Rosenberg, A. F. & Sanz, I. Understanding B-cell activation and autoantibody repertoire selection in systemic lupus erythematosus: a B-cell immunomics approach. *Immunol. Rev.* **284**, 120–131 (2018).
40. Richardson, C. et al. Molecular basis of 9G4 B cell autoreactivity in human systemic lupus erythematosus. *J. Immunol.* **191**, 4926–4939 (2013).
41. Marcovecchio, G. E. et al. Thymic epithelium abnormalities in DiGeorge and Down syndrome patients contribute to dysregulation in T cell development. *Front. Immunol.* **10**, 447 (2019).
42. Pierce, M. J., LaFranchi, S. H. & Pinter, J. D. Characterization of thyroid abnormalities in a large cohort of children with Down syndrome. *Horm. Res. Paediatr.* **87**, 170 (2017).
43. Mader, S., Brimberg, L. & Diamond, B. The role of brain-reactive autoantibodies in brain pathology and cognitive impairment. *Front. Immunol.* **8**, 1101 (2017).
44. Lepagnol-Bestel, A. M. et al. DYRK1A interacts with the REST/NRSF-SWI/SNF chromatin remodelling complex to deregulate gene clusters involved in the neuronal phenotypic traits of Down syndrome. *Hum. Mol. Genet.* **18**, 1405–1414 (2009).
45. van der Poel, C. E., Spaapen, R. M., van de Winkel, J. G. J. & Leusen, J. H. W. Functional characteristics of the high affinity IgG receptor, FcγRI. *J. Immunol.* **186**, 2699–2704 (2011).
46. Bustamante, J., Boisson-Dupuis, S., Abel, L. & Casanova, J. L. Mendelian susceptibility to mycobacterial disease: genetic, immunological, and clinical features of inborn errors of IFN-γ immunity. *Semin. Immunol.* **26**, 454–470 (2014).
47. Malle, L. et al. Individuals with Down syndrome hospitalized with COVID-19 have more severe disease. *Genet. Med.* **23**, 576–580 (2020).
48. Gillespie, K. M. et al. Islet autoimmunity in children with Down's syndrome. *Diabetes* **55**, 3185–3188 (2006).
49. Kohen, D. & Wise, P. H. Autoantibodies in Down's syndrome. *Lancet* **340**, 430 (1992).
50. Cardinale, K. M. et al. Immunotherapy in selected patients with Down syndrome disintegrative disorder. *Dev. Med. Child Neurol.* **61**, 847–851 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2023

Methods

Sample collection

The study was approved under protocols at Mount Sinai Health System (MSHS) (IRB-18-00638/ STUDY-18-00627 and IRB-20-03276), Boston Children's Hospital (04-09-113R), National Institute of Allergy and Infectious Disease (NIAID, NIH) (05-I-0213), Rockefeller University (JCA-0700 and XFK-0815), the French Ethics Committee Comité de Protection des Personnes, the French National Agency for Medicine and Health Product Safety and the Institut National de la Santé et de la Recherche Médicale (protocols C10-13 and C10-14). Patients were approached in person at the hospital or by email obtained through the NIH's DS-Connect national registry (https://dsconnect.nih.gov). All patient data were deidentified. Written informed consent for all individuals in this study was provided in compliance with an institutional review board protocol. All samples from uninfected patients were drawn in the context of an outpatient routine visit and patients exhibited no signs of infection (fever, runny nose, cough, sore throat). Demographics and characteristics of individuals with DS are outlined in Extended Data Tables 1 and 2. For the six patients with SLE included in Fig. 3 and Extended Data Fig. 4, four had severe disease with renal involvement, one had moderate disease without renal involvement and one had mild disease with renal involvement that was resolved at time of sampling (Extended Data Table 3). Half of the patients had active disease and the other three donors were in remission on treatment. From each donor, blood was drawn into a cell preparation tube with sodium heparin (BD Vacutainer). Plasma was isolated from Ficoll separation and stored at −80 °C until use. Whole blood was either directly fixed using Proteomic Stabilizer PROT1 (SmartTube) and frozen at −80 °C, or fresh whole-blood samples were stained by adding 270 µl blood directly to a MDIPA tube containing the lyophilized antibody panel, mixing and incubating for 30 min at room temperature, followed by stabilization with PROT1 and freezing.

Determination of clinical immune dysfunction score

A clinical questionnaire was filled out by patients' legal guardian or referring physician. One point was administered for each autoimmune manifestation or infectious episode. In cases in which an episode (such as pneumonia) was characterized as recurrent or reported more than three times, we capped the score for this category at three points. For example, if a patient had hypothyroidism and 12 ear infections, the total immune clinical score for this patient would be 4 points.

Mass cytometry

The detailed protocol for mass cytometry was previously described[51]. Frozen stabilized blood samples were thawed according to the manufacturer's recommended protocol, then washed with barcode permeabilization buffer (Fluidigm). The samples were uniquely barcoded using the Cell-ID 20-Plex Pd Barcoding Kit (Fluidigm) and pooled together. For previously unstained samples, cells were incubated with an antibody cocktail for surface markers to identify major immune populations, followed by methanol permeabilization, heparin-block and stain with a cocktail of antibodies against intracellular targets, including markers of phosphorylation and signalling. All of the following antibodies were purchased from Fluidigm: Cd111-granzyme B, Cd112-IgA, In113-CD57, In115-CD11c, Cd116-IgD, I127-127I, Ce140-140Ce, Pr141-Ki67, Nd142-CD19, Nd143-CD45RA, Nd144-CD103, Nd145-CD4, Nd146-CD8, Sm147-pSTAT5, 150Nd-pSTAT5, Nd148-CD16, Sm149-CD127, Sm149-pSTAT6, Nd150-CD1c, Eu151-CD123, Sm152-CD66b, Eu153-pSTAT1, Sm154-ICOS, Gd155-CD27, Gd156-p38, 158Gd-pSTAT3, Tb159-pMAPKAP2, Gd160-CD14, Dy161-CD56, Dy162-TCRgd, Dy162-CD169, Dy163-CD172a/b, Dy164-CD69, Ho165-CD64, Ho165-STAT3, Er166-CD25, Er167-pERK1/2, Er168-CD3, Tm169-CD71, Tm169-STAT1, Er170-CD38, Yb171-CD95, Yb171-CD141, Yb172-CD39, Yb173-Tbet, Yb174-HLADR, Lu175-pS6, Yb176-CD54, Pr141-IFNγ, Nd144-CD141, 171Yb-CD141, Sm147-IL-1B, Sm149-IL-1RA, Eu153-TNF, Gd156-IL-6, Gd158-IL-2, Tb159-GM-CSF, Dy164-IL-17A, Ho165-CCL4, Er166-IL-10, Tm169-IFNα2b, Yb173-IL-8, Lu175-IL-29 and Yb176-CXCL10. After washing, cells were incubated in freshly diluted 2.4% formaldehyde containing 125 nM Ir Intercalator (Fluidigm), 0.02% saponin and 30 nM OsO_4 (ACROS Organics) for 30 min at room temperature. Samples were then washed and acquired immediately.

For acquisition, samples were washed with PBS + 0.2% BSA, PBS and then CAS buffer (Fluidigm). The final solution in CAS buffer consisted of 1 million cells per ml and a 1/20 dilution of EQ beads (Fluidigm). After routine instrument optimization, samples were acquired at a rate of <300 events per second on a Helios mass cytometer (Fluidigm) with a modified wide-bore injector (Fluidigm).

FCS files of acquired events were normalized and concatenated with Fluidigm acquisition software, deconvoluted using a MATLAB-based debarcoding application and the resulting files were analysed using Cytobank. Cell events were identified as Ir191/193-positive and Ce140-negative events. Doublets were excluded on the basis of Mahalanobis distance and barcode separation and with the Gaussian parameters calculated using the Helios CyTOF software. Downstream data analysis was performed on Cytobank, by biaxial gating of immune populations according to a previously published gating scheme[51] outlined in Supplementary Fig. 1. All cell frequency analyses were restricted to adult HCs and adults with DS (aged at least 18 years) unless otherwise indicated.

Flow cytometry

B cell phenotyping. Cryopreserved PBMCs were thawed and allowed to rest briefly in complete RPMI medium supplemented with 10% FBS, 1% penicillin–streptomycin and 1% GlutaMax. Cells were immunostained with antibodies in 0.5% BSA in PBS for 1 h, washed 3 times in 0.5% BSA in PBS for 1 h and acquired immediately. The following antibodies were used: CD19 APC-Cy7 (SJ25C1), CD27 FITC (M-T271), CD38 APC (HIT2), CD38 PE-Cy7 (HIT2), CD11c PE (B LY6), IgD BV421 (IA6), CD21 APC (Bu32) and anti-human 9G4 IgG APC (provided by J. Farmer).

pSTAT1 staining. The IgG fraction was isolated from HC and DS plasma using the Protein G HP SpinTrap (Sigma-Aldrich). THP-1 cells were incubated with the IgG fraction diluted 1:3 in complete RPMI with 10% FBS for 20 min or recombinant anti-IFNGR2 antibodies (2 µg ml^{-1}, Thermo Fisher Scientific, PA5-47938) at 37 °C for 20 min followed by stimulation with IFNγ (10 ng ml^{-1}) for 20 min at 37 °C. Cells were stained with Zombie Aqua Live-Dead (BioLegend) for 30 min on ice, then fixed/permeabilized in 90% ice-cold methanol and stained with anti-phospho-STAT1-PE (1:25, BD) for 1 h on ice. Flow cytometry was acquired on the BD LSR Fortessa II system, and data were analysed with FlowJo.

Whole-blood cytokine blocking experiments

Whole blood was incubated for 4 h with tofacitinib (500 nM, ApexBio) or tocilizumab (50 µg ml^{-1}, Selleckchem) or vehicle controls (DMSO and human IgG control, 50 µg ml^{-1}, BioLegend, respectively). For IFN-I blocking, whole blood was incubated with a combination of anti-IFNAR2 (2.5 µg ml^{-1} PBL Assay Science), anti-IFNα (0.2 µg ml^{-1}, PBL 31110−1) and IFNβ (0.2 µg ml^{-1}, PBL 31401-1) antibodies or vehicle control for 4 h. For IL-10 blocking, whole blood was incubated with a combination of anti-IL-10 (5 µg ml^{-1}, BioLegend) and anti-IL-10R (5 µg ml^{-1}, BioLegend) for 4 h. After blocking, whole blood was fixed using Proteomic Stabilizer PROT1 (1.4× blood volume, SmartTube) for 10 min at room temperature and frozen at −80 °C.

B cell stimulation

For plasmablast differentiation of $CD11c^+$ cells (Fig. 4a), $2–5 \times 10^4$ naive ($CD19^+CD38^-CD27^-CD11c^-$) and $CD11c^+$ ($CD19^+CD11c^+$) B cells were sort-purified from the PBMCs of healthy donors using the IMI5L sorter (BD) and rested overnight in complete RPMI. They were then stimulated

with IL-21 (50 ng ml^{-1}; BioLegend), BAFF (100 ng ml^{-1}, BioLegend), IL-10 (250 ng ml^{-1}, BioLegend), IL-2 (5 ng ml^{-1}) and R848 (1 μg ml^{-1}; Invivogen) with or without IFNγ diluted in complete RPMI. Cells were collected at day 4 and stained for flow cytometry.

For the cytokine-induced plasmablast differentiation of B cells (Fig. 4d,e and Extended Data Fig. 5a–f), 2–5 × 10^4 sorted naive or total B cells were sorted from healthy donor PBMCs using fluorescence-activated cell sorting, the Pan B Cell Isolation Kit II or the Naive B Cell Isolation Kit II (Miltenyi Biotech), as indicated. They were cultured for 3–6 days in complete RPMI supplemented with IL-21 (50 ng ml^{-1}; BioLegend), BAFF (10 ng ml^{-1}, BioLegend), IL-2 (50 ng ml^{-1}, BioLegend), R848 (1 μg ml^{-1}; Invivogen) with or without IL-10 (250 ng ml^{-1}, BioLegend) and/or anti-human IgM (5 μg ml^{-1}, Southern Biotech) as indicated, in addition to either (1) plasma from HC individuals or individuals with DS that was depleted from IgG using the Protein G HP SpinTrap (Sigma-Aldrich) at a final dilution of 1:4 or (2) exogenous addition of IL-6 (100 ng ml^{-1}, BioLegend), IFNα2b (100 U ml^{-1}, Merck), IL-4 (20 ng ml^{-1}, BioLegend) and IFNγ (20 ng ml^{-1}, BioLegend). At least two healthy donors and four individuals with DS were tested for each experiment. Cells were collected at indicated days and stained for flow cytometry.

For cytokine-blocking experiments, sorted naive B cells were cultured as described above in the presence of the following blocking antibodies: anti-IFNGR2 (2 μg ml^{-1}, Thermo Fisher Scientific, PA5-47938), anti-IFNAR2 (2.5 μg ml^{-1} PBL Assay Science), anti-IFNα (0.2 μg ml^{-1}, PBL 31110–1) and IFNβ (0.2 μg ml^{-1}, PBL 31401-1), tocilizumab (50 μg ml^{-1}, Selleckchem), adalimumab (2 μg ml^{-1}, Selleckchem) or a combination. Cells were collected at indicated days and stained for flow cytometry.

T cell–B cell co-cultures

For co-cultures of in vitro polarized T cells (Fig. 4f–h and Extended Data Fig. 5g–i). CD4 T cells were purified using MACS (Miltenyi Biotec) from healthy donor PBMCs and cultured in the presence of plate-bound anti-CD3 (OKT3, 10 μg ml^{-1}) and anti-CD28 (CD28.2, 5 μg ml^{-1}) antibodies in complete RPMI alone (unstimulated control) or with (1) IL-2 (50 ng ml^{-1}), IL-12 (5 ng ml^{-1}) and anti-IL-4 (1 μg ml^{-1}, R&D MAB204-SP) (T_H1 conditions); (2) IL-6 (50 ng ml^{-1}) with or without IL-2 (50 ng ml^{-1}); or (3) plasma from HC individuals or individuals with DS (1:2 ratio). After 4 days, T cells were washed and irradiated (5,000 rad) and co-cultured with freshly MACS-sorted naive B cells from the same donor. Cells were co-cultured for 3–6 days in complete RPMI supplemented with IL-21 (50 ng ml^{-1}), IL-2 (50 ng ml^{-1}) and anti-human IgM (5 μg ml^{-1}, Southern Biotech) and collected for flow cytometry. The supernatants were collected for IFNγ ELISA quantification (BioLegend) according to the manufacturer's instructions.

For T cell–B cell co-cultures from HC donors and donors with DS (Fig. 4i,j), CD4 T cells were MACS-purified and co-incubated with sort-purified CD19$^+$CD38$^-$CD27$^-$CD11c$^-$ B cells on the same day. B and T cells were co-cultured for 3 days in complete RPMI supplemented with IL-21, IL-2 and anti-human IgM as above and collected for flow cytometry.

BCR analysis

Sample preparation. Donor PBMCs were thawed in complete RPMI and rested overnight. Naive (CD19$^+$CD27$^-$CD38^{low}CD11c$^-$), CD11c$^+$ (CD1 9$^+$CD27$^-$CD38^{low}CD11c$^-$) and memory (CD19$^+$CD27$^+$CD38$^-$) B cells were sort-purified using an IMI5L sorter (BD) and genomic DNA was isolated with QIAamp DNA Blood Mini Kit (Qiagen).

Library preparation. Sample data were generated using the immunoSEQ Assay (Adaptive Biotechnologies). The somatically rearranged *Homo sapiens* BCR IgH locus *CDR3* was amplified from genomic DNA using a two-step, amplification-bias-controlled multiplex PCR approach[52,53]. The first PCR consists of forward and reverse amplification primers specific for every V and J gene segment and amplifies the hypervariable complementarity-determining region 3 (*CDR3*) of the immune receptor locus. Furthermore, primers are included to amplify B cell precursor DJ rearrangements. The second PCR adds a proprietary barcode sequence and Illumina adapter sequences[54]. *CDR3* libraries were sequenced on the Illumina instrument according to the manufacturer's instructions.

Data analysis. Raw sequencing reads were demultiplexed according to Adaptive's proprietary barcode sequences. Demultiplexed reads were then processed to remove adapter and primer sequences, identify and remove primer dimer, germline and other contaminant sequences. The filtered data are clustered using the relative frequency ratio between similar clones and a modified nearest-neighbour algorithm to merge closely related sequences to correct technical errors introduced through PCR and sequencing while allowing for somatic hypermutation in the V segment. The resulting sequences were sufficient to allow annotation of the V, D and J genes, and the N1 and N2 regions constituting each unique *CDR3* and the translation of the encoded CDR3 amino acid sequence. Gene definitions were based on annotation in accordance with the IMGT database (www.imgt.org). The set of observed biological BCR IgH CDR3 sequences was normalized to correct for residual multiplex PCR amplification bias and quantified against a set of synthetic BCR IgH CDR3 sequence analogues[53]. Data were analysed using the immunoSEQ Analyzer toolset.

Multiplex cytokine analysis

Plasma collected by Ficoll isolation from heparinized whole blood was clarified by centrifugation. Magnetic Luminex assays with the MILLIPLEX MAP Human Cytokine/Chemokine Magnetic Bead Panel (MilliporeSigma, HCYTMAG-60 K-PX30) were run according to the manufacturer's protocol. The samples were quantified on the MAGPIX xMAP Instrument (Luminex). For each sample, >50 beads were collected per analyte. The median fluorescence intensity of these beads was recorded and used for analysis with the Milliplex Analyst software using a 5P regression algorithm. In the absence of signal, the assay's lower limit of detection for each cytokine was used. Patients with DS aged 7 months to 37 years ($n = 21$) and control individuals aged 6 years to 39 years ($n = 10$) were included.

Auto-antibody analysis

Seromic profiling of autoantibodies was conducted as previously described[55]. The CDI HuProt peptide array was run according to the manufacturer's instructions, at 1/500 to avoid low-tittered cross-reactivity and using a robust blocking buffer to prevent unspecific binding. After applying patient sera, chips were washed and secondary fluorescent antibodies to quantify IgG (Cy3; 532 nm) and IgA (Cy5; 635 nm) antibody isotypes were applied at the manufacturer's recommended concentrations. Each spot on the array, representing a protein printed in duplicate, was gated using Genepix software alignment and then manually quality controlled to ensure proper quantification and removal of possible artifacts. High differences between replicate spots (CV > 0.5) were flagged, along with information about staining artifacts (rare and excluded). Raw sample-peptide matrices were then read into the R (v.4.0.4) statistical environment to be processed by the limma microarray analysis suite (v.3.46.0)[56]. On a per-sample and a per-isotype (IgG or IgA) basis, the background intensity was subtracted using the backgroundCorrect function (method: normexp, normexp.method: rma) and then normalized across all samples using the normalizeBetweenArrays function (method="cyclicloess"). Next, all control and non-detected peptides were filtered from the sample–peptide matrices and duplicate spot intensities were averaged across all of the remaining peptides on a per-sample basis. Individually processed sample matrices were concatenated, column-wise, to generate an experimental matrix containing all samples and peptides shared across all arrays and to be analysed for differential peptide reactivity

using the limma differential expression functions. First, to exclude low-reactive peptides and outlier reactivity, peptides were filtered to include only peptides that had reactivities above the median signal intensity in at least 5 samples (of the 50 total samples processed). PCA was then performed using the base R (v.4.0.4) prcomp function (scale: TRUE) and PCA and loadings were visualized using the fviz_pca_ind function from the FactoMiner/factoextra package (v.1.0.7)[57]. To determine differentially abundant autoantibodies, we first fit a linear model for each peptide to model the variation of peptide reactivity (against an auto-antibody) across all arrays using limma's lmFit function. After building the model, we generated a contrast matrix to test disease groups against healthy controls. Differential autoantibody abundance testing was performed and moderated under the empirical Bayes framework described by the developer and resulting peptide lists were filtered against an absolute $\log_2$[fold change] > 1 and a $P < 0.05$. Overlap analyses were conducted and visualized using the UpSetR (v.1.4.0)[58] and ggvennDiagram (https://github.com/gaospecial/ggVennDiagram) (v.1.2.1) packages. Heat maps were generated using the ComplexHeatmap (v.2.7.4)[59] library.

Chromosomal reactivity heatmap. To assess whether autoantibody reactivity was influenced by the gene dosage effect in DS, filtered differentially abundant autoantibodies (obtained from the DS versus HC contrast; see Supplementary Data 1) were plotted across the 23 human chromosomes (hg38) in a location-dependent manner. In brief, gene names were extracted from their associated peptides and chromosomal locations were identified using the org.Hs.eg.db (v.3.12.0)[60] database. Hits were further filtered against a fold-change of 1.5 for plotting purposes and the karyoplotR (v.1.16.0)[61] package was then used for visualization.

Tissue-protein expression heat map. Tissue-protein detection values of all human proteins were extracted from the Human Protein Atlas API using the HPAanalyze (v.1.8.1)[62] package. Differentially abundant autoantibodies (obtained from the DS versus HC contrast; Supplementary Data 1) were subset from the full dataset (Supplementary Data 2) and visualized using the ComplexHeatmap library.

GO analysis. GO enrichment analysis was conducted on over-abundant ($\log_2$[fold change] > 1) autoantibodies with limma's goana function, and biological process GO terms with $P < 0.0001$ were subset and visualized using the ggplot2[63] (v.3.3.3) package.

ELISA assays

Recombinant protein ELISA. Overnight, 96-well Costar plates were coated at 4 °C with 100 µl per well of a 1 µg ml^{-1} solution of recombinant IFNGR2, MSTN or ATP6V1G2 protein (OriGene) suspended in 1× PBS. The next morning, the coating solution was removed and wells were washed three times with 100 µl of washing buffer (PBS with 0.05% (v/v) Tween 20; PBS-T). Next, 200 µl of blocking buffer (PBS with 1% with bovine plasma albumin (endotoxin-free)) was added to each well at room temperature and incubated at room temperature for 1 h. Plasma samples were diluted 1:100 in blocking buffer.

Total IgG and 9G4 IgG ELISA. Overnight, 96-well Costar plates were coated at 4 °C with 100 µl per well of a 2 µg ml^{-1} solution of goat anti-human IgG, F(ab′)$_2$-fragment-specific (Jackson ImmunoResearch) or 1:100 dilution of anti-human 9G4 IgG antibody (provided by J. Farmer) suspended in 1× PBS. The next morning, the coating solution was removed and wells were washed with three times with 100 µl of washing buffer (PBS with 0.05% (v/v) Tween-20; PBS-T). Next, 200 µl of blocking buffer (PBS with 1% with bovine plasma albumin (endotoxin-free)) was added to each well at room temperature and incubated at room temperature for 1 h. The supernatants from B cell stimulation experiments were diluted 1:10.

Antibody detection. Plates were then washed three times before adding goat anti-human IgG F(ab)–horseradish peroxidase (HRP)-conjugated secondary antibody (diluted 1:3,000 in PBS-T). The plates were incubated for 1 h at 37 °C and washed three times. The plates were again washed three times with wash buffer. Once completely dry, 100 µl SIGMAFAST OPD (*o*-phenylenediamine dihydrochloride; Sigma–Aldrich) solution was added to each well. This substrate was left on the plates for 5 min and the reaction was then stopped by the addition of 50 µl per well of 3 M hydrochloric acid. The optical density at 490 nm was measured using the POLARstar Omega (BMG Labtech) plate reader.

Detection of anti-type I IFN autoantibodies

Detection of anti-type I IFN autoantibodies was performed using Gyros as described previously[64]. Cytokines, recombinant human IFNα2 (Miltenyi Biotec, 130-108-984) or recombinant human IFNω (Merck, SRP3061) were first biotinylated with EZ-Link Sulfo-NHS-LC-Biotin (Thermo Fisher Scientific, A39257), according to the manufacturer's instructions, with a biotin-to-protein molar ratio of 1:12. The detection reagent contained a secondary antibody (Alexa Fluor 647 goat anti-human IgG (Thermo Fisher Scientific, A21445)) diluted in Rexxip F (Gyros Protein Technologies, P0004825; 1:500 dilution of the 2 mg ml^{-1} stock to yield a final concentration of 4 µg ml^{-1}). Phosphate-buffered saline, 0.01% Tween-20 (PBS-T) and Gyros Wash buffer (Gyros Protein Technologies, P0020087) were prepared according to the manufacturer's instructions. Plasma or serum samples were then diluted 1:100 in 0.01% PBS-T and tested with the Bioaffy 1000 CD (Gyros Protein Technologies, P0004253) and the Gyrolab xPand (Gyros Protein Technologies, P0020520). Cleaning cycles were performed in 20% ethanol.

Functional evaluation of anti-type I IFNs autoantibodies using luciferase reporter assays

The blocking activity of anti-IFNα2 and anti-IFNω autoantibodies was determined with a reporter luciferase activity as described previously[64]. HEK293T cells were transfected with a plasmid containing the firefly luciferase gene under the control of the human ISRE promoter in the pGL4.45 backbone and a plasmid constitutively expressing *Renilla* luciferase for normalization (pRL-SV40). Cells were transfected in the presence of the X-tremeGENE9 transfection reagent (Sigma-Aldrich, 6365779001) for 24 h. Cells in Dulbecco's modified Eagle medium (DMEM; Thermo Fisher Scientific) supplemented with 2% fetal calf serum and 10% healthy control or patient serum/plasma (after inactivation at 56 °C for 20 min) were stimulated with IFNα2 (Miltenyi Biotec, 130-108-984) and IFNω (Merck, SRP3061) at 10 ng ml^{-1} or 100 pg ml^{-1} for 16 h at 37 °C. Each sample was tested once for each cytokine and dose. Finally, cells were lysed for 20 min at room temperature, and luciferase levels were measured with the Dual-Luciferase Reporter 1000 Assay System (Promega, E1980) according to the manufacturer's protocol. Luminescence intensity was measured with a VICTOR-X Multilabel Plate Reader (PerkinElmer Life Sciences). Firefly luciferase activity values were normalized against *Renilla* luciferase activity values.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data supporting the findings of this study are available in the Article and its Supplementary Information.

51. Geanon D, et al. A streamlined whole blood CyTOF workflow defines a circulating immune cell signature of COVID-19. *Cytometry A.* **99**, 446-461 (2021).
52. Robins, H. S. et al. Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells. *Blood* **114**, 4099–4107 (2009).

53. Carlson, C. S. et al. Using synthetic templates to design an unbiased multiplex PCR assay. *Nat. Commun.* **4**, 2680 (2013).
54. Robins, H. et al. Ultra-sensitive detection of rare T cell clones. *J. Immunol. Methods* **375**, 14–19 (2012).
55. Gnjatic, S. et al. Seromic profiling of ovarian and pancreatic cancer. *Proc. Natl Acad. Sci. USA* **107**, 5088–5093 (2010).
56. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
57. Lê, S., Josse, J. & Husson, F. FactoMineR: an R package for multivariate analysis. *J. Stat. Softw.* **25**, 1–18 (2008).
58. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).
59. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
60. Carlson, M. org.Hs.eg.db (2017).
61. Gel, B. karyoploteR (2017).
62. Nhat, A. HPAanalyze (2018).
63. Wickham, H. ggplot2 (2009); https://doi.org/10.1007/978-0-387-98141-3.
64. Bastard, P. et al. Autoantibodies neutralizing type I IFNs are present in ~4% of uninfected individuals over 70 years old and account for ~20% of COVID-19 deaths. *Sci. Immunol.* **6**, abl4340 (2021).

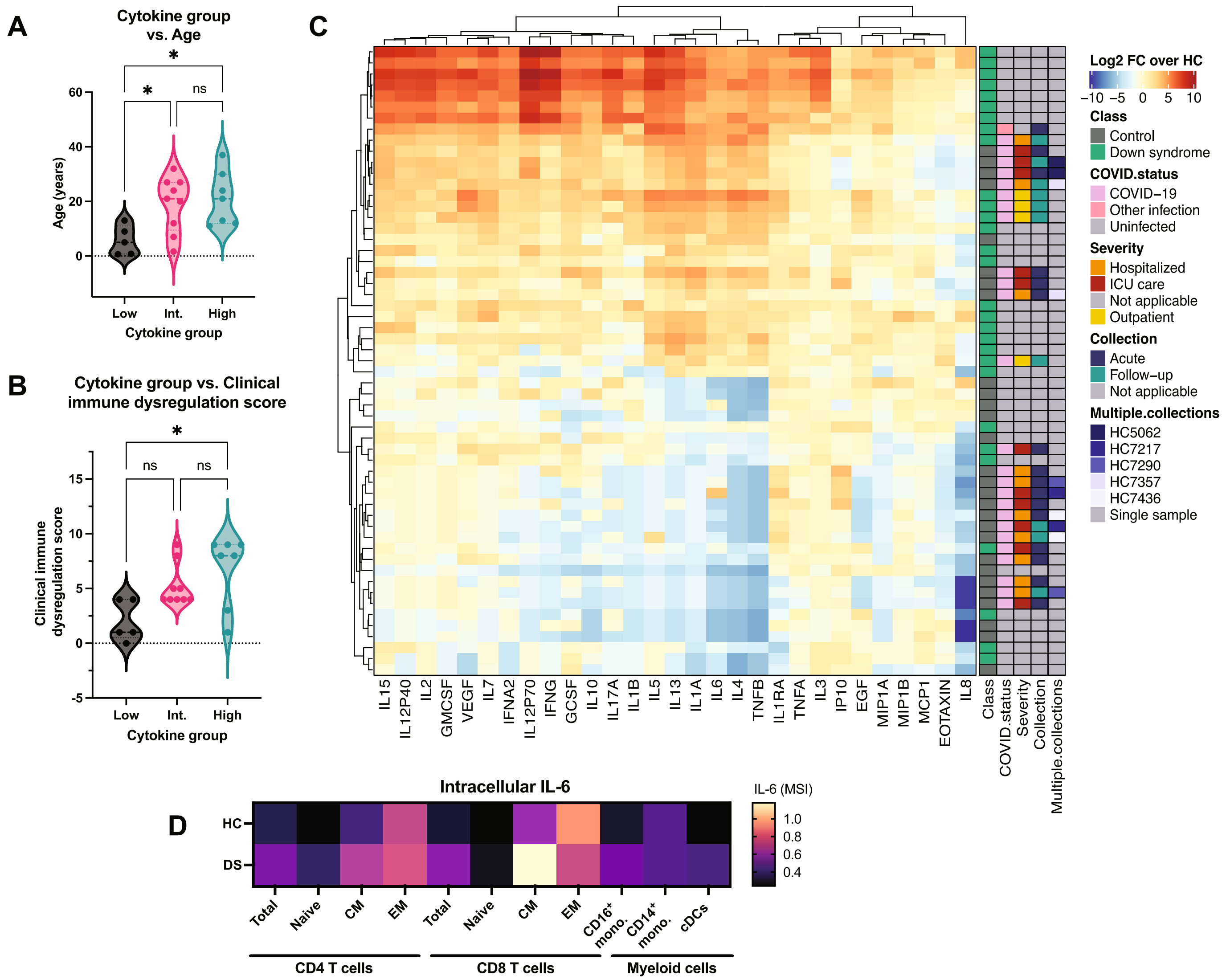
Acknowledgements We thank all of the patients and their families for their participation; A. Rahman, D. Geanon and G. Kelly from the Human Immune Monitoring Centre at the Icahn School of Medicine for their technical assistance; and J. Farmer and K. Hillier for sharing reagents. This study was funded by the National Institute of Allergy and Infectious Diseases Grants R01AI150300, R01AI150300-01S1 and R01AI151029. L. Malle was supported by the National Institute of Child Health and Human Development T32 training grant T32HD075735. L.N. is supported by the Division of Intramural Research, National Institute of Allergy and Infectious Diseases, NIH. S.G. was supported by NIH grants CA224319, DK124165 and CA196521.

Author contributions L. Malle designed and performed experiments and analysed the data for cytokine array, CyTOF, B–T cell co-cultures and BCR sequencing, and wrote the manuscript. R.S.P. analysed the CDI array and edited the manuscript. M.M.-F. performed steady-state B–T cell co-cultures from DS donors. O.S. performed 9G4 and IFNγ ELISA. J.T. edited the manuscript. Q.P. performed IFN autoantibody Gyros and neutralization assays. S.B. and A.R. coordinated DS cohort and processed whole-blood samples. V.B., K.T. and S.G. performed the CDI array. B.R.R. helped to analyse BCR sequencing data. P.B., J.S., C.M., A.-S.R., L. Maillebouis, M.V.-M., R.T., J.-L.C., L.N. and D. Bush recruited patients. D. Bogunovic supervised the work, wrote the manuscript, and helped to design the experiments and analyse the data.

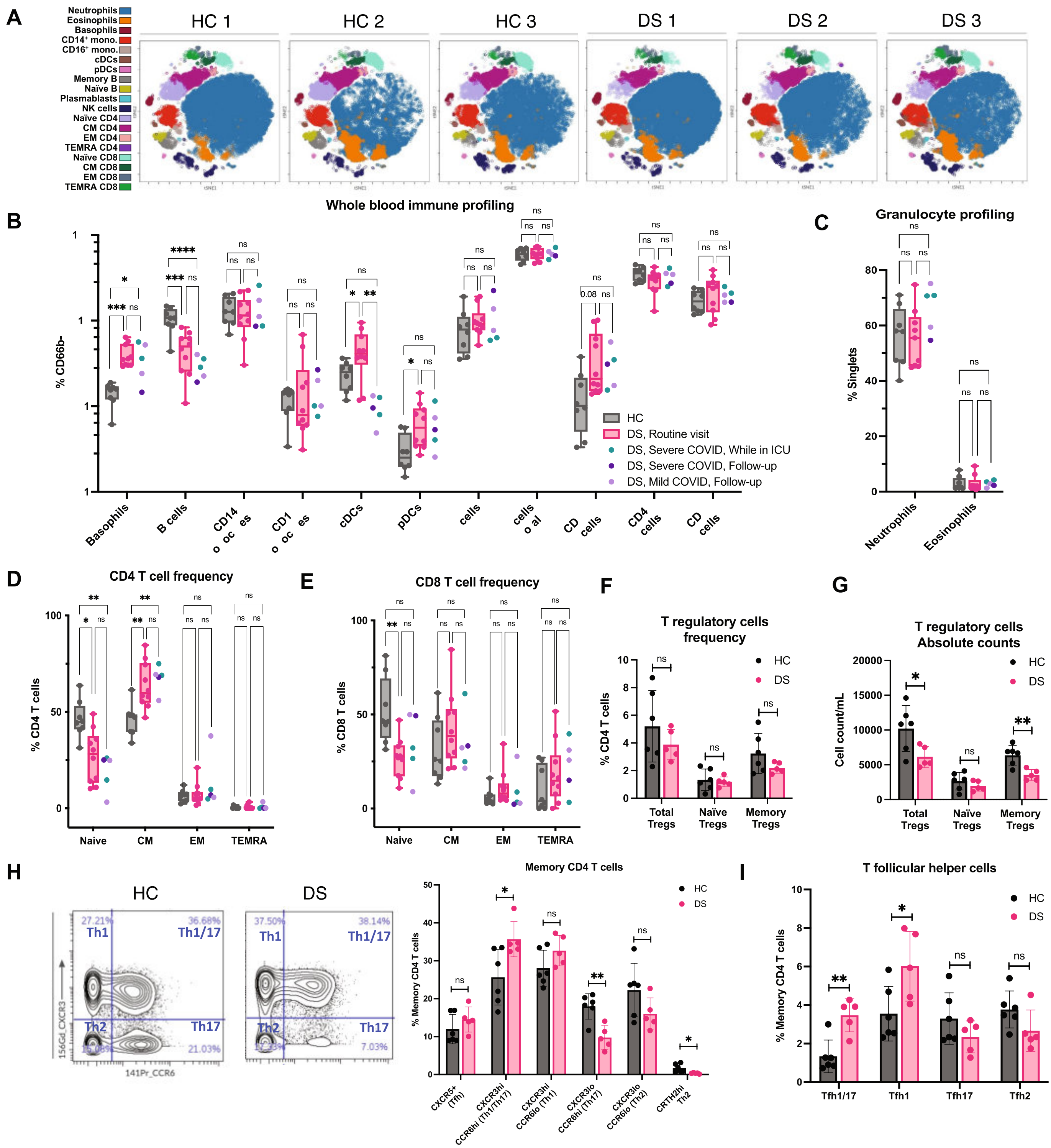
Competing interests D.B. is the founder and part owner of Lab11 Therapeutics. S.G. reports other research funding from Genentech, Boehringer-Ingelheim, Celgene, Takeda, and Regeneron.

Additional information

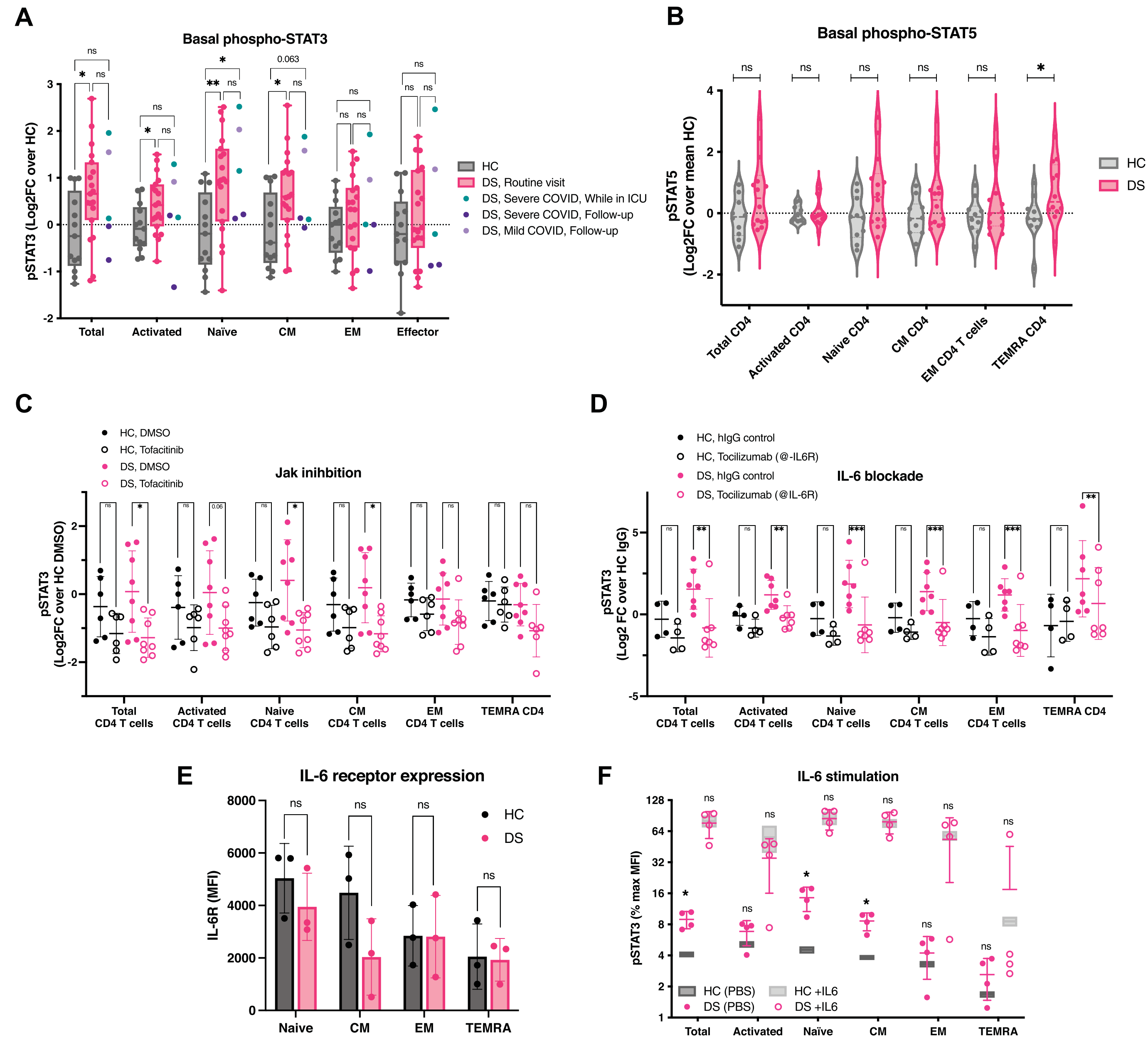
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-05736-y.
Correspondence and requests for materials should be addressed to Dusan Bogunovic.
Peer review information *Nature* thanks Stuart Tangye and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



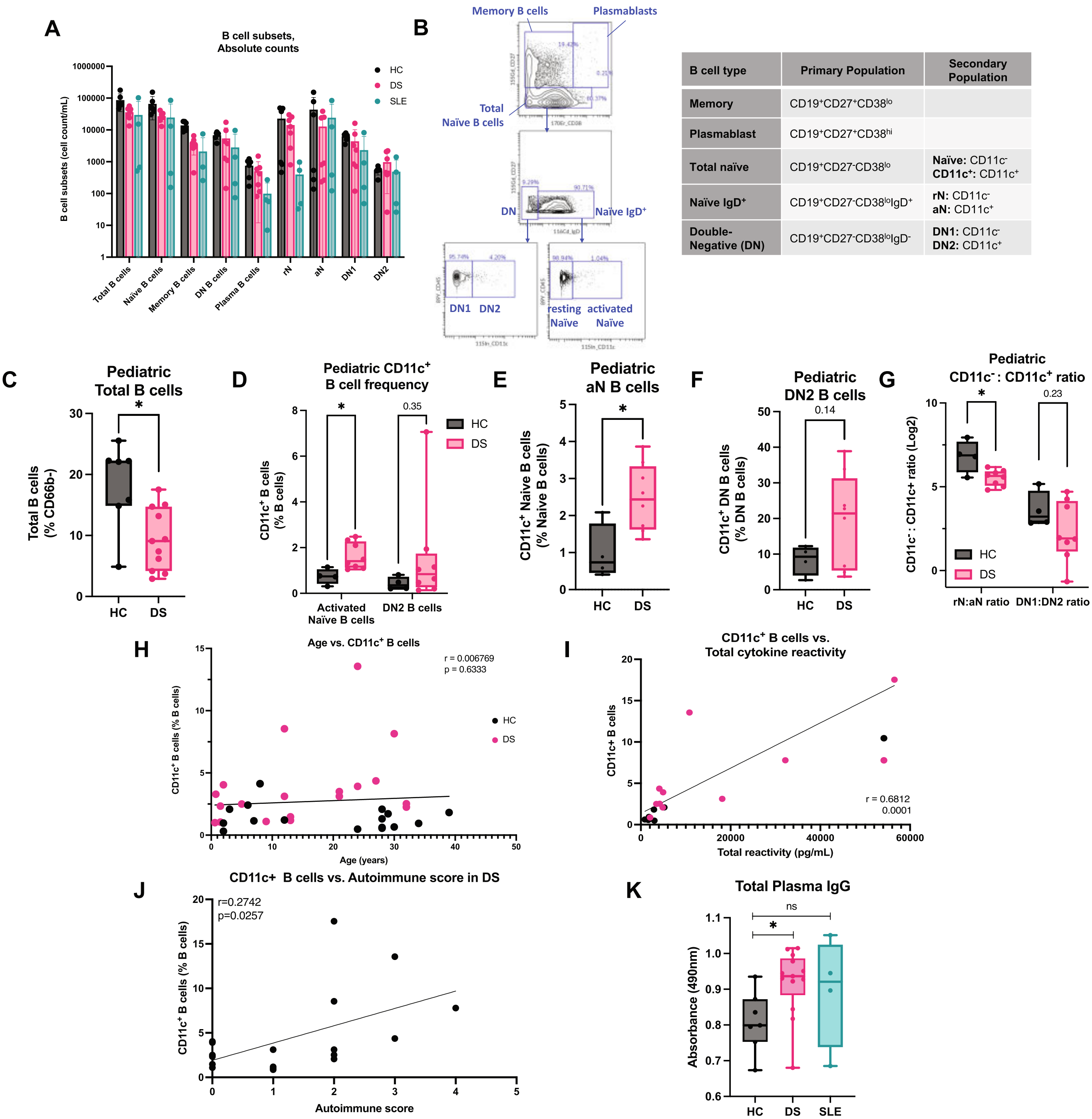
Extended Data Fig. 1 | Global cytokine dysregulation in DS. (**A-B**) Correlation between cytokine group and (**A**) Age and (**B**) Clinical Immune Dysfunction score (determined according to reported clinical history, see Methods) and Age in individuals with DS. Significance assessed by one-way ANOVA with Tukey's post-hoc analysis, ns denotes $p > 0.05$, *$p \leq 0.05$; **$p \leq 0.005$, ***$p \leq 0.0005$. (**C**) Multiplex cytokine analysis by Magnetic Luminex assay of plasma from HCs (n = 10), uninfected individuals with DS (n = 21), and individual with acute respiratory infection (n = 1), or individuals with DS and COVID-19 (n = 7) with severity and time of sampling as indicated, expressed as log2FC over the mean HC per cytokine. Unsupervised clustering of samples and cytokines using the complete method (distance metric: Euclidean). (**D**) Intracellular staining of IL-6 in CD4 T cells, CD8 T cells, and Myeloid cells from HCs (n = 4) and individuals with DS (n = 5).



Extended Data Fig. 2 | Cellular immune landscape in DS. (**A**) Representative t-SNE of agranulocytes adults with DS (n = 3) and age-matched HCs (n = 3) illustrating the immune cell distribution in whole blood. (**B-E**) Frequencies of (**B**) agranulocyte subsets, (**C**) granulocyte subsets, and (**D**) CD4 and (**E**) CD8 T cell subsets in whole blood from adults with DS (n = 10), adults with DS and COVID-19 (n = 5), and age-matched HCs (n = 8). (**F-G**) Frequencies (**F**) and absolute counts (**G**) of T regulatory (Treg) cells ($CD4^{+}CD25^{+}CD127^{-}$) and naive ($CD45RA^{+}$) and memory ($CD45RA^{-}$) subsets in whole blood from adults with DS (n = 5) and age-matched HCs (n = 5). (**H-I**) Representative plots and calculated frequencies of (**H**) T helper (Th) and (**I**) T follicular helper (Tfh) cell subsets in whole blood from adults with DS (*n* = 5) and age-matched HCs (*n* = 5), expressed as percent of memory CD4 T cells. In (B-E), whiskers denote min and max values, bounds of box denote Q1–Q3, and centre bar denotes mean. (F-I) Error bars denote SD. (B-E) Significance assessed by one-way ANOVA with Tukey's post-hoc analysis, ns denotes $p > 0.05$, *$p \leq 0.05$; **$p \leq 0.005$, ***$p \leq 0.0005$. (F-I) Significance assessed by two-tailed paired t tests, ns denotes $p > 0.05$, *$p \leq 0.05$; **$p \leq 0.005$, ***$p \leq 0.0005$.

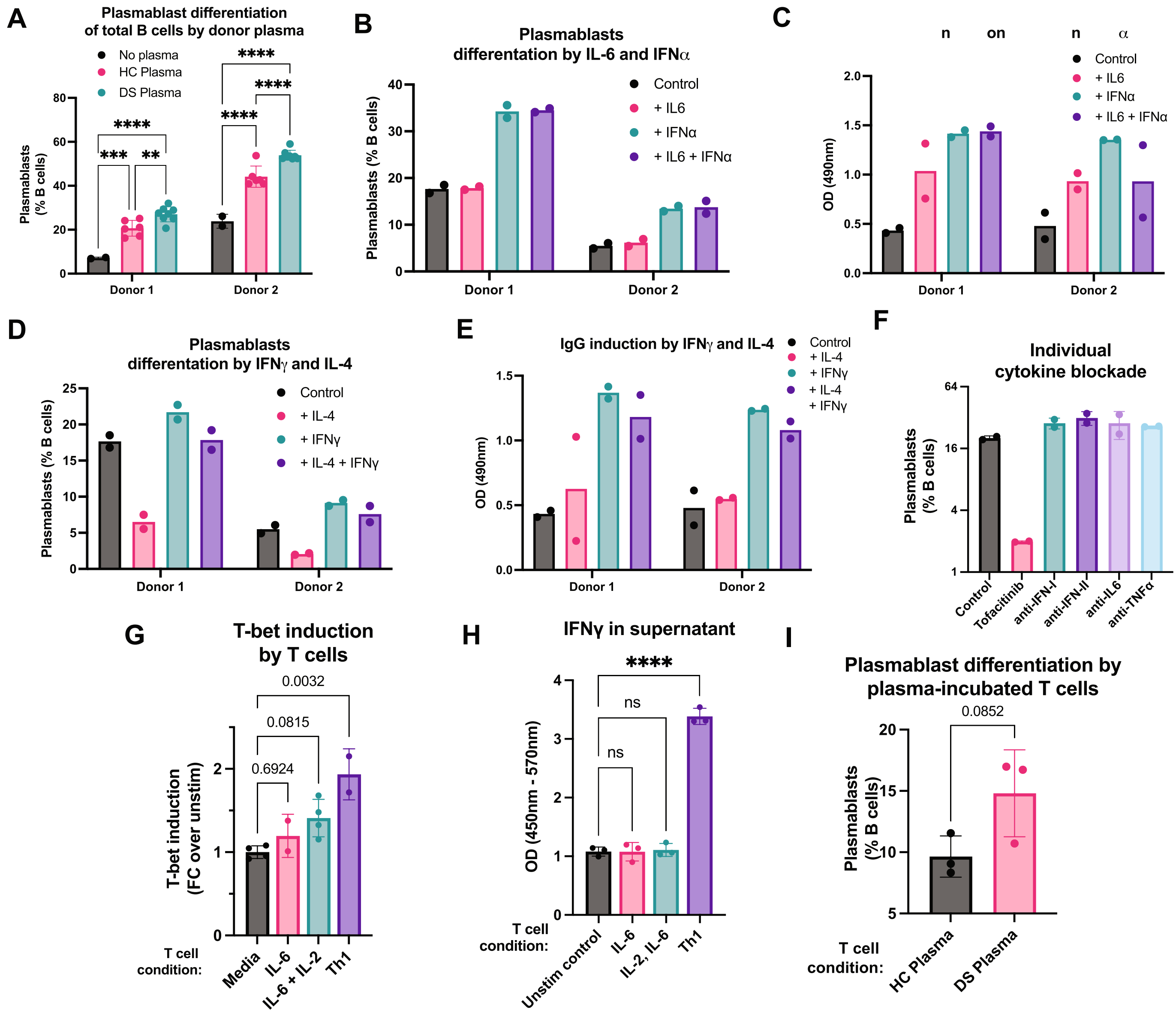


Extended Data Fig. 3 | Basal signalling in CD4 T cells in DS. (**A**) Basal STAT3 phosphorylation in CD4 T cell subsets from HCs (n = 13), individuals with DS (n = 19), and adults with DS and COVID-19 (n = 5) expressed as Log2FC over the mean HCs per subset. Significance assessed by one-way ANOVA with Tukey's post-hoc analysis, ns denotes p > 0.05, *p ≤ 0.05; **p ≤ 0.005, ***p ≤ 0.0005. Whiskers denote min and max values, bounds of box denote Q1–Q3, and centre bar denotes mean. (**B**) Basal STAT5 phosphorylation in CD4 T cell subsets from individuals with DS (n = 14) and age-matched controls (n = 10), expressed as Log2FC over the mean HCs per subset. (**C**-**D**) STAT3 phosphorylation in CD4 T cell subsets expressed as Log2FC over the mean mock-treated HCs per subset after ex vivo whole blood treatment for 4 h with (**C**) Tofacitinib (500nM) (n = 6 HC, n = 7 DS), (**D**) Tocilizumab (50 $\mu g\ ml^{-1}$) (n = 4 HC, n = 7 DS). (**E**) Surface expression of IL-6R in CD4 T cells from adults with DS (n = 3) and age-matched HCs (n = 3). (**F**) STAT3 phosphorylation in CD4 T cell subsets induced by stimulation of whole blood with recombinant IL-6 (50 $ng\ ml^{-1}$) for 15 min, expressed as Log2FC over the mean mock-treated HCs (n = 2 HC, n = 4 DS). Box plots denote min and max values for HCs, error bars indicate SD and centre denotes mean for DS. (C-E) Error bars denote SD. (B-F) Significance assessed by two-tailed paired t tests, ns denotes p > 0.05, *p ≤ 0.05; **p ≤ 0.005, ***p ≤ 0.0005.

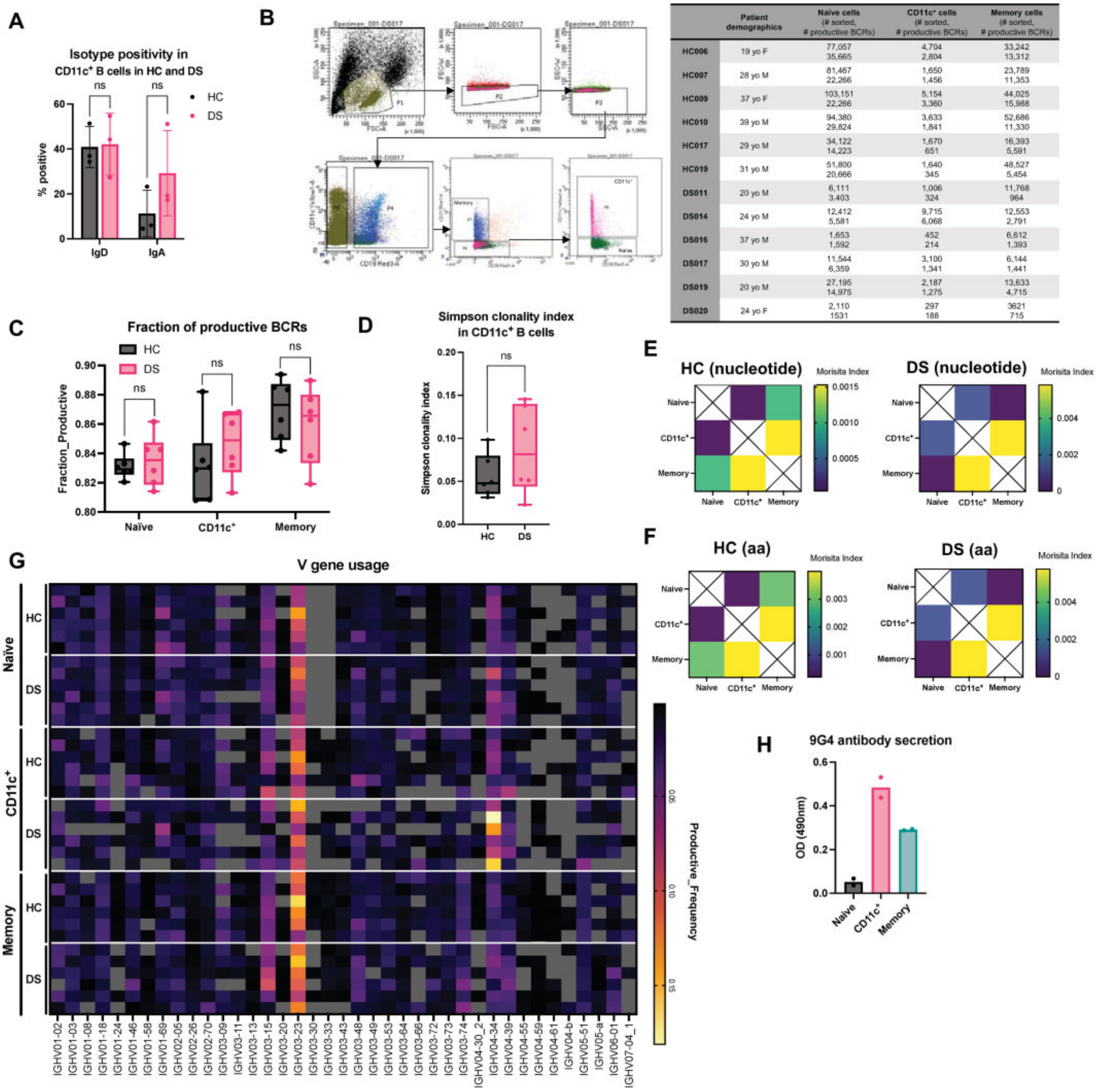


B cell type	Primary Population	Secondary Population
Memory	CD19+CD27+CD38lo	
Plasmablast	CD19+CD27+CD38hi	
Total naïve	CD19+CD27-CD38lo	Naïve: CD11c- CD11c+: CD11c+
Naïve IgD+	CD19+CD27-CD38loIgD+	rN: CD11c- aN: CD11c+
Double-Negative (DN)	CD19+CD27-CD38loIgD-	DN1: CD11c- DN2: CD11c+

Extended Data Fig. 4 | Abnormal B cell subsets in DS. (**A**) Raw B cell counts in HCs (n = 6) and adults with DS (n = 7) or patients with SLE (n = 4). Error bars denote SD. (**B**) Gating scheme of B cells subtypes. (**C**) Frequency in children with DS (n = 11) and age-matched HCs (n = 7) of total B cells expressed as percent of CD66b⁻ cells (non-granulocytes), Significance assessed by two-tailed unpaired t tests, *p ≤ 0.05. (**D**-**F**) Frequency in children with DS (n = 8) and age-matched HCs (n = 4) of CD11c⁺ B cells in the IgD⁺ naive or DN compartments, expressed as percent of (**D**) total, (**E**) IgD⁺ naïve, or (**F**) DN B cells. (**G**) Ratios of CD11c⁺ subsets to CD11c⁻ subset (rN:aN and DN2:DN1) in HC and DS groups, Log2-transformed. Significance assessed by two-tailed unpaired t tests, *p ≤ 0.05. (**H**-**I**) Correlation of CD11c⁺ B cell frequency and (**H**) age and (**I**) total cytokines (calculated as the sum of all circulating cytokines, pg ml^{-1}) in individuals (H) or adults (I) with DS and age-matched controls. (r: Pearson correlation coefficient). (**J**) Correlation of CD11c⁺ B cell frequency and autoimmune score (calculated as the sum of autoimmune diseases) in individuals with DS. (r: Pearson correlation coefficient). (**K**) Total IgG in the plasma of HCs (n = 7) and individuals with DS (n = 12) or SLE (n = 6) assessed by ELISA. Significance assessed by one-way ANOVA with Tukey's post-hoc analysis, ns denotes p > 0.05, *p ≤ 0.05. In (C-F, K), whiskers denote min and max values, bounds of box denote Q1–Q3, and centre bar denotes mean.

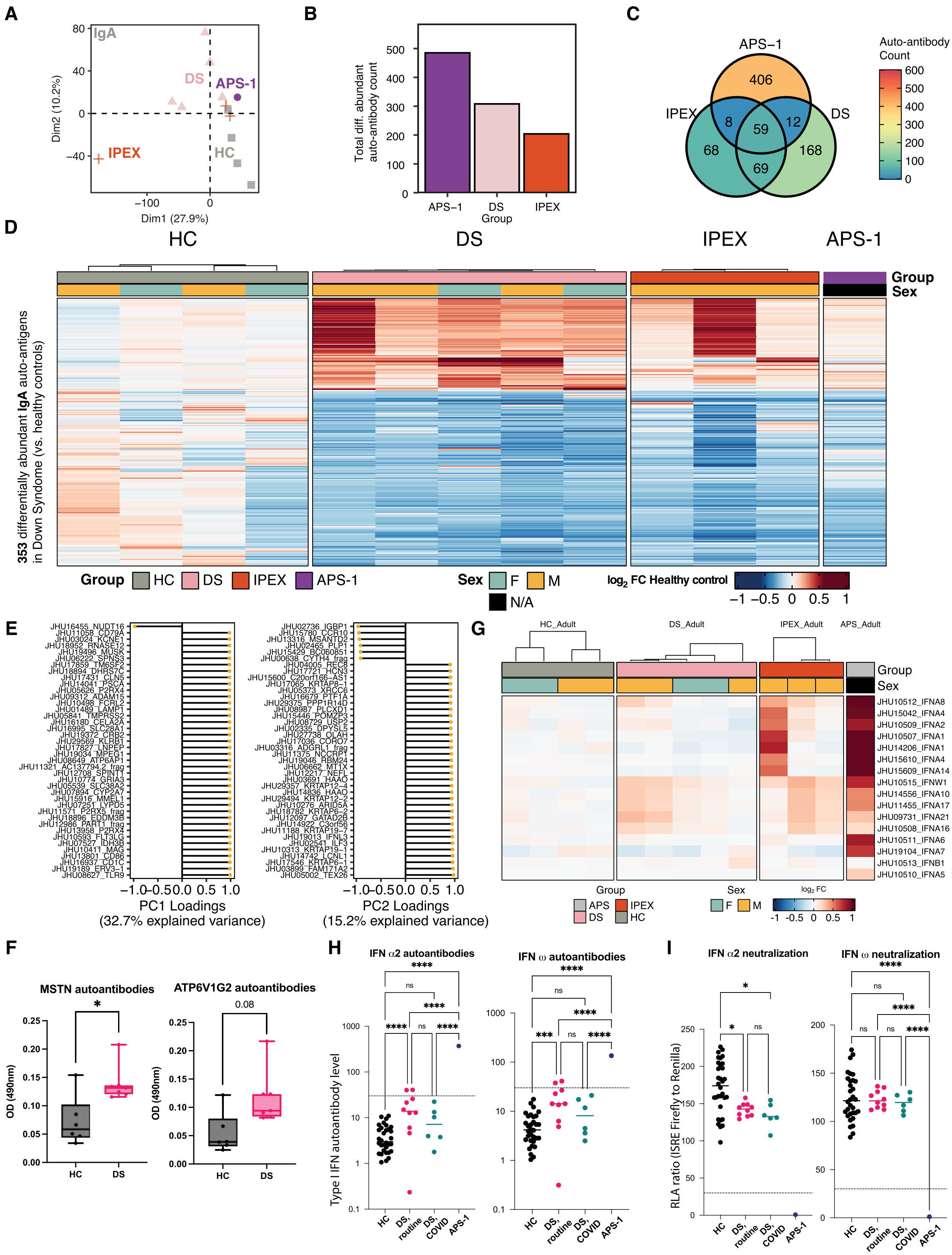


Extended Data Fig. 5 | Atypical activation of naive B cells by cytokines and activated T cells. (**A** Plasmablast differentiation of MACS-isolated total B cells from 2 healthy donors after 3-day culture in the presence of BAFF (10 ng ml^{-1}), IL-2 (50 ng ml^{-1}), IL-21 (50 ng ml^{-1}), R848 (1 μg ml^{-1}), anti-IgM (1 μg ml^{-1}), and IgG-depleted plasma from HCs (n = 3) or individuals with DS (n = 3), run in duplicates. (**B**-**C**) MACS-isolated total B cells from 2 healthy donors were cultured in the presence of BAFF, IL-2, IL-21, R848, anti-IgM (same concentrations as A), in the presence of IL-6 (100 ng ml^{-1}), IFN-α2b (100 U ml^{-1}), or both, run in duplicates. After 3 days, cells were washed and cultured for another 3 days in media with anti-IgM. (**B**) Plasmablast differentiation at day 3 and (**C**) ELISA for total IgG in supernatant at day 6. (**D**-**E**) MACS-isolated total B cells from 2 healthy donors were cultured for 3-days in the presence of BAFF, IL-2, IL-21, R848, anti-IgM (same concentrations as A), and IL-4 (20 ng ml^{-1}), IFN-γ (20 ng ml^{-1}), or both, run in duplicates. Cells were then washed and cultured for another 3 days in media with anti-IgM. (**D**) Plasmablast differentiation at day 3 and (**E**) ELISA for total IgG in supernatant at day 6. (**F**) Plasmablast differentiation of MACS-isolated naive B cells from a healthy donor after 3-day culture in the presence of BAFF, IL-2, IL-21, R848, anti-IgM (same concentrations as J), and IgG-depleted DS plasma in the presence of Tofacitinib (500nM) or antibodies blocking IFN-I, IFN-II, IL-6, and TFN-α signalling, run in duplicates. (**G**-**H**) Co-cultures containing T cells activated with IL-6, IL-2, both, or polarized into "Th1 cells" together with MACS-isolated naive B cells from the same donor, run in triplicates. (**G**) Intracellular T-bet expression in non-plasmablast B cells and (**H**) quantification of IFN-g in the supernatant after 3-6 days of co-culture. (**I**) Frequency of plasmablasts in co-cultures containing T cells previously polarized with serum from HCs (n = 3) or individuals with DS (n = 3) together with MACS-isolated naive B cells from the same donor. Significance assessed by two-tailed unpaired t-test. (A-I) Error bars denote SD. (A,H) Significance assessed by One-way ANOVA with Tukey's post-hoc analysis, ns denotes $p > 0.05$; $^{**}p \leq 0.005$; $^{***}p \leq 0.0005$; $^{****}p \leq 0.0001$.



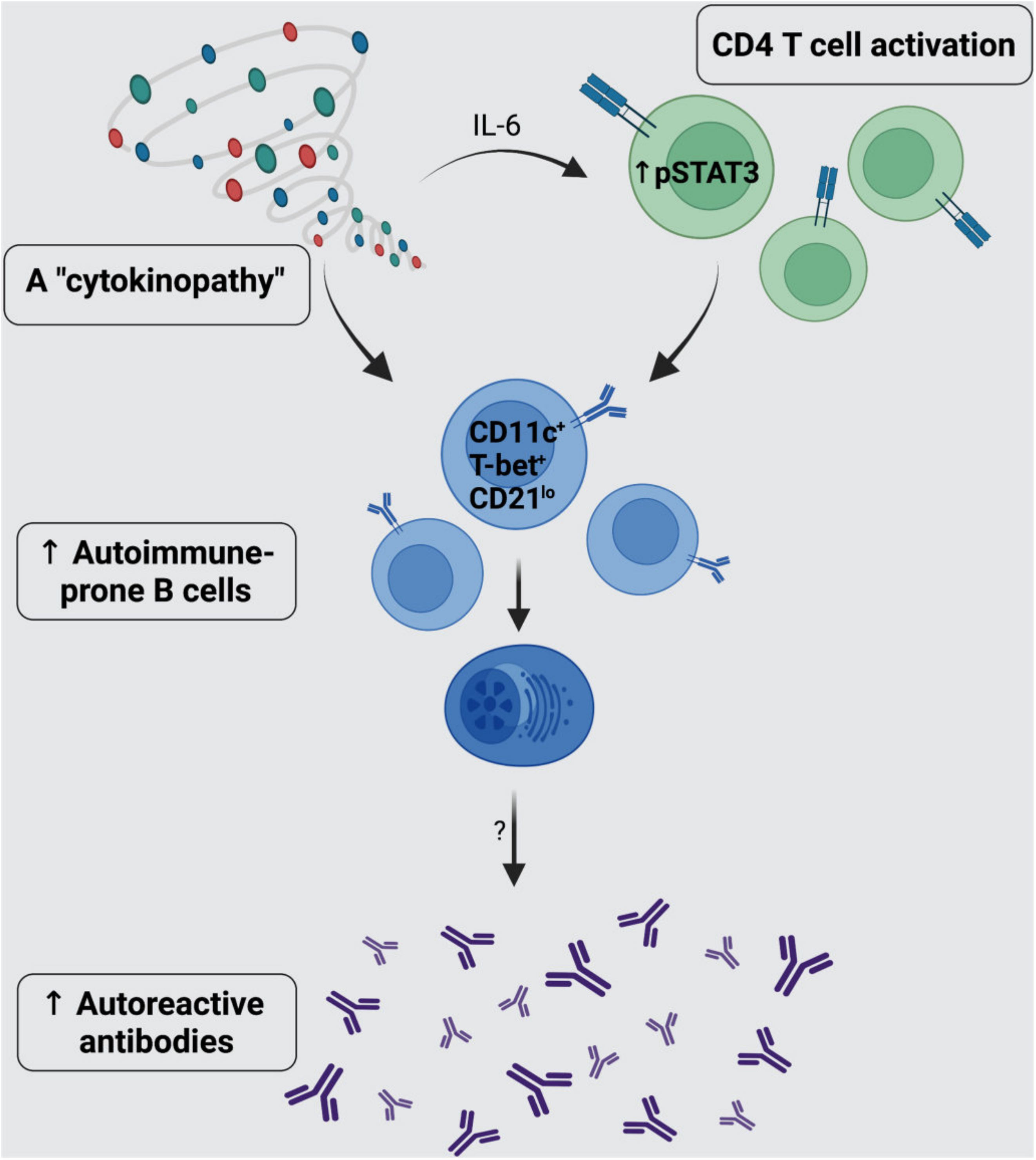
	Patient demographics	Naïve cells (# sorted, # productive BCRs)	CD11c⁺ cells (# sorted, # productive BCRs)	Memory cells (# sorted, # productive BCRs)
HC006	19 yo F	77,057 35,665	4,704 2,804	33,242 13,312
HC007	28 yo M	81,467 22,266	1,650 1,456	23,789 11,353
HC009	37 yo F	103,151 22,266	5,154 3,360	44,025 15,988
HC010	39 yo M	94,380 29,824	3,633 1,841	52,686 11,330
HC017	29 yo M	34,122 14,223	1,670 651	16,393 5,591
HC019	31 yo M	51,800 20,666	1,640 345	48,527 5,454
DS011	20 yo M	6,111 3,403	1,006 324	11,768 964
DS014	24 yo M	12,412 5,581	9,715 6,068	12,553 2,791
DS016	37 yo M	1,653 1,592	452 214	6,612 1,393
DS017	30 yo M	11,544 6,359	3,100 1,341	6,144 1,441
DS019	20 yo M	27,195 14,975	2,187 1,275	13,633 4,715
DS020	24 yo F	2,110 1531	297 188	3621 715

Extended Data Fig. 6 | Receptor sequencing in atypical B cells. (**A**) Expression of IgD and IgA in CD11c⁺ B cells from adults with DS (n = 3) and age-matched HCs (n = 3). Significance assessed by unpaired t-tests, ns denotes $p > 0.05$. Error bars denote SD. (**B-F**) BCR sequencing from gDNA isolated from sorted naïve, CD11c⁺ and memory B cells from controls (n = 6) and individuals with DS (n = 6). (**B**) Sorting scheme for naïve, CD11c⁺ and memory B cells (left) and number of cells sorted and number of productive BCRs obtained by sequencing for each cell type (right). (**C**) Fraction of in-frame BCRs containing no stop codons ("productive BCRs") in naïve, CD11c⁺ and memory B cells from controls (n = 6) and individuals with DS (n = 6). (**D**) Simpson clonality in productive BCRs from CD11c⁺ B cells from controls (n = 6) and individuals with DS (n = 6). In (C-D), significance assessed by two-tailed unpaired t-test (ns denotes $p > 0.05$) and whiskers denote min and max values, bounds of box denote Q1–Q3, and centre bar denotes mean. (**E**-**F**) Representative heatmaps of Morisita index indicating overlap between B cell subsets in HC and DS at the (**E**) nucleotide and (**F**) amino acid levels. (**G**) Heatmap of IGHV gene frequency in HC and DS. Only genes that occurred at higher than 0.1% frequency in more than 10 samples are shown. (**H**) 9G4 IgG antibodies in supernatant after 4-day culture of sorted HC naïve, CD11c⁺ or memory B cells in the presence of BAFF, IL-2, IL-10, IL-21, the TLR7/8 ligand R848 with or without IFN-γ. Results representative of 2 independent experiments.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Autoantibodies in DS plasma. (**A**) Principal component analysis of HuProt IgA dataset for adult HCs (n = 4), adults with DS (n = 5), and patients, Immunodysregulation polyendocrinopathy enteropathy X-linked syndrome (IPEX) (n = 3) and Autoimmune polyglandular syndrome type 1 (APS-1) (n = 1). (**B**) Number of IgA autoantigens enriched at least 2-fold in DS, IPEX, and APS-1 compared to HCs. (**C**) Venn diagram of enriched IgA autoantigens overlapping between disease groups. (**D**) Heatmap of enriched IgA autoantigens in HCs, DS, IPEX, and APS-1. Colour intensity corresponds to the log2FC expression value relative to the mean of healthy adult controls. (**E**) Component loadings (PC1 and PC2) from Principal component analysis of HuProt IgG dataset for adults with DS (n = 5) and age-matched HCs (n = 4). (**F**) ELISA of anti-MSTN and anti-ATP6V1G2 autoantibodies in plasma from HCs (n = 6) and individuals with DS (n = 6). Significance assessed by two-tailed unpaired t tests, * denotes $p \leq 0.05$. Whiskers denote min and max values, bounds of box denote Q1–Q3, and centre bar denotes mean. (**G**) Heatmap of IgG type I IFN autoantigens in HCs, DS, IPEX, and APS-1. Colour intensity corresponds to the log2FC expression value relative to the mean of healthy adult controls. (**H**) Quantification of antibodies against IFNα2 and IFNω in serum from HCs (n = 32), DS without COVID-19 (n = 10), DS with COVID-19 (n = 6), and APS-1 (n = 1) by Gyros assay. (**I**) Neutralization of IFNα2 and IFNω activity by serum from HCs (n = 32), DS without COVID-19 (n = 10), DS with COVID-19 (n = 6), and APS-1 (n = 1) by Gyros assay. (H-I) Significance assessed by One-way ANOVA with Tukey's post-hoc analysis, ns denotes $p > 0.05$, *$p \leq 0.05$; **$p \leq 0.005$; ***$p \leq 0.0005$; ****$p \leq 0.0001$.



Extended Data Fig. 8 | Graphical abstract. Breaking Immune Tolerance in Down Syndrome: A Triad of Cytokines, Activated T cells and CD11c+ B Cells. Created with BioRender.

Extended Data Table 1 | Clinical history and calculated immune dysfunction scores of individuals with DS

	Sex	Age at 1st draw	RSV	Respiratory infections (non-RSV) 1=1pt, 2=2pts, 3 or recurrent=3pts	Ear infections 1=1pt, 2=2pts, 3 or recurrent=3pts	Other infections:	Skin manifestations 1=1pt, 2=2pts, 3 or recurrent=3pts	Celiac disease (1pt)	Hypo-thyroidism (1pt)	Type 1 diabetes mellitus (1pt)	Asthma	Allergies	Other immune condition	Auto-immune score	TOTAL CLINICAL IMMUNE DYSFUNCTION SCORE
DS001	F	7 yo	0	3	0	0	0	0	0	0	0	1	0	1	**4**
DS002	M	9 months	1	0	0	0	0	0	1	0	0	0	0	1	**2**
DS003	F	7 months	0	0	0	0	0	0	1	0	0	0	0	1	**1**
DS005	F	11 yo	0	0	0	0	0	1	0	0	0	0	0	1	**1**
DS008	M	4 yo	0	3	0	0	0	0	0	0	0	0	0	0	**3**
DS009	M	12 yo	1	0	1	0	0	0	1	0	0	1	0	2	**4**
DS013	M		0	3	3	0	0	0	1	0	0	1	0	2	**8**
DS015	M	20 mo	1	3	0	0	0	0	0	0	0	0	0	0	**4**
DS021	M	19 mo	0	0	0	0	0	0	0	0	0	0	0	0	**0**
DS022	F	13 yo	0	3	0	0	0	0	1	0	0	0	0	1	**4**
DS023	F	9 yo	1	3	0	0	0	0	0	0	0	0	0	0	**4**
DS024	F	13 yo	0	0	2	1 (meningitis)	0	0	0	0	0	0	0	0	**3**
DS028	F	2 yo	0	0	0	0	0	0	0	0	0	0	0	0	**0**
DS006	F	32	0	1	3	2 (fungal infections)	1 (skin boils)	0	1	0	0	1	0	2	**9**
DS010	M	27	0	1	1	3 (Scarlett fever, fungal infection, MRSA)	1 (skin boils)	0	1	0	0	1	0	2	**8**
DS011	M	20	2	0	3	3 (Sinus infections)	0	0	1	0	0	0	0	1	**9**
DS012	M	27	0	0	0	0	1 (Liver abscess)	1	1	1	0	0	0	3	**4**
DS014	M	24	0	0	3	3 (Sinus infections, hangnail)	1 (chronic folliculitis)	1	1	0	0	0	0	3	**9**
DS016	M	37	0	3	3	0		0	1	0	0	1	0	2	**8**
DS017	M	30	0	0	3	1 (conjuctivitis)	1 (elastosis perforans serpiginosa)	0	1	1	0	1	1 (anti-phospholipid syndrome)	4	**9**
DS018	M	20	0	3	0	0	0	0	1	0	0	0	0	1	**4**
DS019	M	20	1	1	3	0	0	0	0	0	0	0	0	0	**5**
DS020	F	24	0	2	3	0	0	0	0	0	0	0	0	0	**5**

Extended Data Table 2 | Clinical summary of the individuals with DS studied

Category	Number
n	23
Mean age (Range, SD), years	15.3 (0.6-38, 11.2)
Gender Male (n, %)	14 (60.9%)
Gender Female (n, %)	9 (39.1%)
RSV (n, %)	6 (27.2%)
1 or more respiratory infections (n, %)	13 (56.5%)
Repeated (>2) respiratory infections (n, %)	9 (43.4%)
Repeated (>2) ear infections (n, %)	7 (30.4%)
Celiac disease (n, %)	2 (8.7%)
Hypothyroidism (n, %)	12 (52.2%)
Diabetes (n, %)	2 (8.7%)
Type I (n, %)	2 (8.7%)
Type II (n, %)	0
Allergies or asthma (n, %)	8 (34.7%)
Rxn to vaccination (n, %)	0

Extended Data Table 3 | Immunological and clinical features of individuals with SLE studied

Patient ID	PBs (% B cells)	aN (% B cells)	DN2 (% B cells)	rN:aN (Log2)	DN1:DN2 (Log2)	Age	Gender	Severity of disease	Duration of disease	Renal involve-ment	Active disease or remission at time of blood draw	Current Treatments
SLE001	1.4493	11.2069	35	2.98606031	0.8930848	18	F	Severe	2 years	Yes	Yes	Prednisone, hydroxychloroquine, tacrolimus, mycophenolate mofetil, lisinopril, rituximab
SLE002	0.4717	4.3912	21.9048	4.44445626	1.83398684	15	F	Severe	~3 years	Yes	No	labetalol, amlodopine, prednisone, mycophenolate sodium, ondansetron ODT, darbepoetin alfa-polysorbate, hydroxychloroquine
SLE003	0.2649	1.1503	14.5881	6.42515452	2.54312735	17	M	Mild	7 years	in past, resolved years ago	No	hydrochloroquine
SLE004	8.2873	21.4286	23.6364	1.87446667	1.6918748	20	F	Severe	3.5 years	Yes	Yes	Rituximab, tacrolimus, enalapril, hydrochloroquine, furosemide, prednisone, mycophenolate mofetil
SLE005	0.0473	1.979	9.0358	5.63024735	3.32772184	12	F	Severe	1 year	Yes	No	hydroxychloroquine, tacrolimus, lisinopril, mycophenolate mofetil, prednisone
SLE006	0.4151	4.1228	33.7766	4.5394914	0.96551042	19	F	Moderate	1.5 years	No	Yes	celecoxib, gabapentin, mycophenolate mofetil, lhydroxychloroquine

nature portfolio

Corresponding author(s): Dusan Bogunovic

Last updated by author(s): 12/14/2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	CyTOF FCS files of acquired events were normalized and concatenated with Fluidigm acquisition software. Flow cytomoetry was acquired with FACSDiva software (BD Bioscience). Luminex data was acquired with Milliplex Analyst software using a 5P regression algorithm.
Data analysis	Autoantibody analysis was done in R (v4.0.4) using the following packages: limma microarray analysis suite (v3.46.0), UpSetR (v1.4.0)59 and ggvennDiagram (https://github.com/gaospecial/ggVennDiagram) (v1.2.1) packages. Heatmaps were generated using the ComplexHeatmap (v2.7.4), HPAanalyze (v1.8.1). CyTOF analysis was performed in Cytobank. Flow cytometry analysis was performed with FlowJo. BCR sequencing analysis was performed usign the immunoSEQ Analyzer toolset. Graphpad Prism 9 and Microsoft Excel 16.57 were also used for data analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data that support the findings of this study are available in the article and supplementary materials

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	The study cohort was comprised of both sexes as shown in Supplemental Tables 1 and 2. Sexes were included whenever possible. There were no significant differences between genders for any of the reported findings.
Population characteristics	N/A
Recruitment	"Individuals with DS" participants with a diagnosis of Down syndrome were recruited, via their referring physicians or via the NIH's DS-Connect ® national registry (dsconnect.nih.gov).
Ethics oversight	Mount Sinai Health System (MSHS) (IRB-18-00638/ STUDY-18-00627 and IRB-20-03276), Boston Children's Hospital (04-09-113R), National Institute of Allergy and Infectious Disease (NIAID, NIH) (05-I-0213), Rockefeller University (JCA-0700 and XFK-0815), the French Ethics Committee "Comité de Protection des Personnes," the French National Agency for Medicine and Health Product Safety, and the "Institut National de la Santé et de la Recherche Médicale" (protocols # C10-13 and C10-14).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	DS (n=23) and HC (n=21)
Data exclusions	N/A
Replication	Experiments were repeated with a minimum of n=2 in each condition
Randomization	Individuals were randomly allocated within HC and DS groups
Blinding	Blinding was not performed in our study

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Cd111-GranzymeB, Cd112-IgA, In113-CD57, In115-CD11c, Cd116-IgD, I127-127I, Ce140-140Ce, Pr141-Ki67, Nd142-CD19, Nd143-CD45RA, Nd144-CD103, Nd145-CD4, Nd146-CD8, Sm147-pSTAT5, 150Nd-pSTAT5, Nd148-CD16, Sm149-CD127 Sm149-pSTAT6, Nd150-CD1c, Eu151-CD123, Sm152-CD66b, Eu153-pSTAT1, Sm154-ICOS, Gd155-CD27, Gd156-p38, 158Gd-pSTAT3, Tb159-pMAPKAP2, Gd160-CD14, Dy161-CD56, Dy162-TCRgd, Dy162-CD169, Dy163-CD172a_b, Dy164-CD69, Ho165-CD64, Ho165-STAT3, Er166-CD25, Er167-pERK1_2, Er168-CD3, Tm169-CD71, Tm169-STAT1, Er170-CD38, Yb171-CD95, Yb171-CD141, Yb172-CD39, Yb173-Tbet, Yb174-HLADR, Lu175-pS6, Yb176-CD54, Pr141-IFNg, Nd144-CD141, 171Yb-CD141, Sm147-IL_1b, Sm149-IL_1RA, Eu153-TNFa, Gd156-IL_6, Gd158-IL_2, Tb159-GM_CSF, Dy164-IL_17A, Ho165-CCL4, Er166-IL_10, Tm169-IFNa2b, Yb173-IL_8, Lu175-IL_29, Yb176-CXCL10.
CD19 APC-Cy7 (SJ25C1), CD27 FITC (M-T271), CD38 APC (HIT2), CD38 PE-Cy7 (HIT2), CD11c PE (B LY6), IgD BV421 (IA6), CD21 APC (Bu32), anti-phospho-STAT1-PE (1:25, BD) , anti-human 9G4 IgG APC (generously provided by Jocelyn Farmer).
Tofacitinib (500nM, ApexBio), Tocilizumab (50ug/mL, Selleckchem), anti-IFNAR2 (2.5ug/mL PBL Assay Science), anti-IFN-a (0.2ug/mL, PBL 31110–1), and IFN-b (0.2ug/mL, PBL 31401-1), anti-IL10 (5ug/mL, Biolegend), anti-IL-10R (5ug/mL, Biolegend), nti-IFNGR2 (2ug/mL, Thermofisher PA5-47938), Adalimumab (2ug/mL, Selleckchem).

Validation

N/A

Clinical data

Policy information about clinical studies
All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration

N/A

Study protocol

Mount Sinai Health System (MSHS) (IRB-18-00638/ STUDY-18-00627 and IRB-20-03276), Boston Children’s Hospital (04-09-113R), National Institute of Allergy and Infectious Disease (NIAID, NIH) (05-I-0213), Rockefeller University (JCA-0700 and XFK-0815), the French Ethics Committee “Comité de Protection des Personnes,” the French National Agency for Medicine and Health Product Safety, and the “Institut National de la Santé et de la Recherche Médicale” (protocols # C10-13 and C10-14).

Data collection

N/A

Outcomes

N/A

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

CYTOF: Frozen stabilized blood samples were thawed according to the manufacturer’s recommended protocol, then washed with barcode permeabilization buffer (Fluidigm). Samples were uniquely barcoded with Cell-ID 20-Plex Pd Barcoding Kit (Fluidigm) and pooled together. For previously unstained samples, cells were then incubated with an antibody cocktail for surface markers to identify major immune populations, followed by methanol permeabilization, heparin-block and stain with a cocktail of antibodies against intracellular targets, including markers of phosphorylation and signaling. After washing, cells were then incubated in freshly diluted 2.4% formaldehyde containing 125nM Ir Intercalator (Fluidigm), 0.02% saponin and 30 nM OsO4 (ACROS Organics) for 30 min at room temperature. Samples were then washed and acquired immediately.
Flow: For extracellular markers, cells were immunostained with antibodies in 0.5% BSA in PBS for 1 hour, washed 3x in 0.5% BSA in PBS for 1 hour and acquired immediately.

Instrument	CyTOF: Helios mass cytometer (Fluidigm) with a modified wide-bore injector (Fluidigm). Flow: BD LSR Fortessa II
Software	CyTOF: Fluidigm acquisition software Flow: BD FACSDiva software
Cell population abundance	See figures
Gating strategy	Major populations were The gated populations were manually gated based on the previously described gating scheme (Geanon, D. et al. A Streamlined CyTOF Workflow To Facilitate Standardized Multi-Site Immune Profiling of COVID-19 Patients. medRxiv (2020)). B cell populations were gated according to gating in Supplemental Figure 4B.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Article

TET2 guards against unchecked BATF3-induced CAR T cell expansion

https://doi.org/10.1038/s41586-022-05692-z

Received: 21 October 2021

Accepted: 30 December 2022

Published online: 8 February 2023

Check for updates

Nayan Jain[1,2,5], Zeguo Zhao[2,5], Judith Feucht[2,4], Richard Koche[3], Archana Iyer[2], Anton Dobrin[1,2], Jorge Mansilla-Soto[2], Julie Yang[3], Yingqian Zhan[3], Michael Lopez[2], Gertrude Gunset[2] & Michel Sadelain[2]✉

Further advances in cell engineering are needed to increase the efficacy of chimeric antigen receptor (CAR) and other T cell-based therapies[1–5]. As T cell differentiation and functional states are associated with distinct epigenetic profiles[6,7], we hypothesized that epigenetic programming may provide a means to improve CAR T cell performance. Targeting the gene that encodes the epigenetic regulator ten–eleven translocation 2 (*TET2*)[8] presents an interesting opportunity as its loss may enhance T cell memory[9,10], albeit not cause malignancy[9,11,12]. Here we show that disruption of *TET2* enhances T cell-mediated tumour rejection in leukaemia and prostate cancer models. However, loss of TET2 also enables antigen-independent CAR T cell clonal expansions that may eventually result in prominent systemic tissue infiltration. These clonal proliferations require biallelic *TET2* disruption and sustained expression of the AP-1 factor BATF3 to drive a MYC-dependent proliferative program. This proliferative state is associated with reduced effector function that differs from both canonical T cell memory[13,14] and exhaustion[15,16] states, and is prone to the acquisition of secondary somatic mutations, establishing TET2 as a guardian against BATF3-induced CAR T cell proliferation and ensuing genomic instability. Our findings illustrate the potential of epigenetic programming to enhance T cell immunity but highlight the risk of unleashing unchecked proliferative responses.

CARs are synthetic receptors for antigens that instruct T cell specificity and augment antitumour functions[2,4]. CAR T cell therapy for relapsed and refractory acute lymphoblastic leukaemia, non-Hodgkin lymphoma and multiple myeloma yields a high rate of complete responses, although a large fraction of patients will eventually relapse from their disease[3,17]. Novel strategies are needed to augment the overall efficacy of CAR T cells to prevent these relapses and tackle solid tumour therapy[2,3,18]. We hypothesized that epigenome programming could act in concert with CARs to promote CAR T cell activity by supporting T cell proliferation and functional persistence. TET2 is a member of the TET family of epigenetic regulators that successively oxidize 5-methyl cytosine in DNA[19]. A study of the T cell receptor in transgenic mice[9] and a case report of a patient with lymphoma with a hypomorphic *TET2* allele treated with CAR T cells[10] suggest that loss of TET2 may enhance T cell responses. Mutations in *TET2* are frequent in myeloid and lymphoid malignancies but are not sufficient to establish a malignant state[20–22]. Here we report unexpected antigen-independent clonal expansions of CAR T cells lacking TET2, which is dependent on sustained expression of BATF3.

Effect of TET2 on CAR T cell efficacy

To assess the effect of TET2 on CAR T cell efficacy, we disrupted *TET2* in human T cells before retrovirally transducing them with either FDA-approved CD28 or 4-1BB-based CD19 CARs, hereafter designated Rv-1928z and Rv-19BBz, respectively, and compared their activity in the well-established B cell acute lymphoblastic leukaemia NALM6 model in NSG mice (Fig. 1a). Human peripheral blood T cells typically showed a CRISPR–Cas9-mediated *TET2* editing efficiency of approximately 67% (Fig. 1b) and retroviral CAR transduction efficiency on the order of approximately 50%. No discernible phenotypic differences were observed between infused edited and control CAR T cells (Extended Data Fig. 1a,b). CAR T cells were administered at low doses to better compare their antitumour efficacy ('stress test' condition[23]). *TET2*-edited Rv-19BBz CAR T cells afforded greater survival of tumour-bearing mice than their unedited counterparts (Fig. 1d and Extended Data Fig. 1d). By contrast, no survival difference was observed between recipients of *TET2*-edited or unedited Rv-1928z CAR T cells (Fig. 1c and Extended Data Fig. 1c). Flow cytometric quantification and phenotyping of CAR T cells isolated from bone marrow and the spleen 3 weeks after their infusion revealed no significant difference in quantity (Extended Data Fig. 1e,f) and differentiation state (Extended Data Fig. 1g,h) between Rv-1928z CAR T cells. However, *TET2*-edited Rv-19BBz CAR T cells were more abundant than their unedited counterparts (Extended Data Fig. 1e,f). *TET2*-edited CAR T cells showed increased expression of CCR7 in Rv-19BBz CAR T cells but not in Rv-1928z CAR T cells (Extended Data Fig. 1g,h), whereas inhibitory receptor expression (PD1, LAG3

[1]Louis V. Gerstner Jr Graduate School of Biomedical Sciences, Memorial Sloan Kettering Cancer Centre, New York, NY, USA. [2]Centre for Cell Engineering and Immunology Program, Memorial Sloan Kettering Cancer Centre, New York, NY, USA. [3]Centre for Epigenetics Research, Memorial Sloan Kettering Cancer Centre, New York, NY, USA. [4]Present address: University Children's Hospital, Tübingen, Germany. [5]These authors contributed equally: Nayan Jain, Zeguo Zhao. ✉e-mail: m-sadelain@ski.mskcc.org

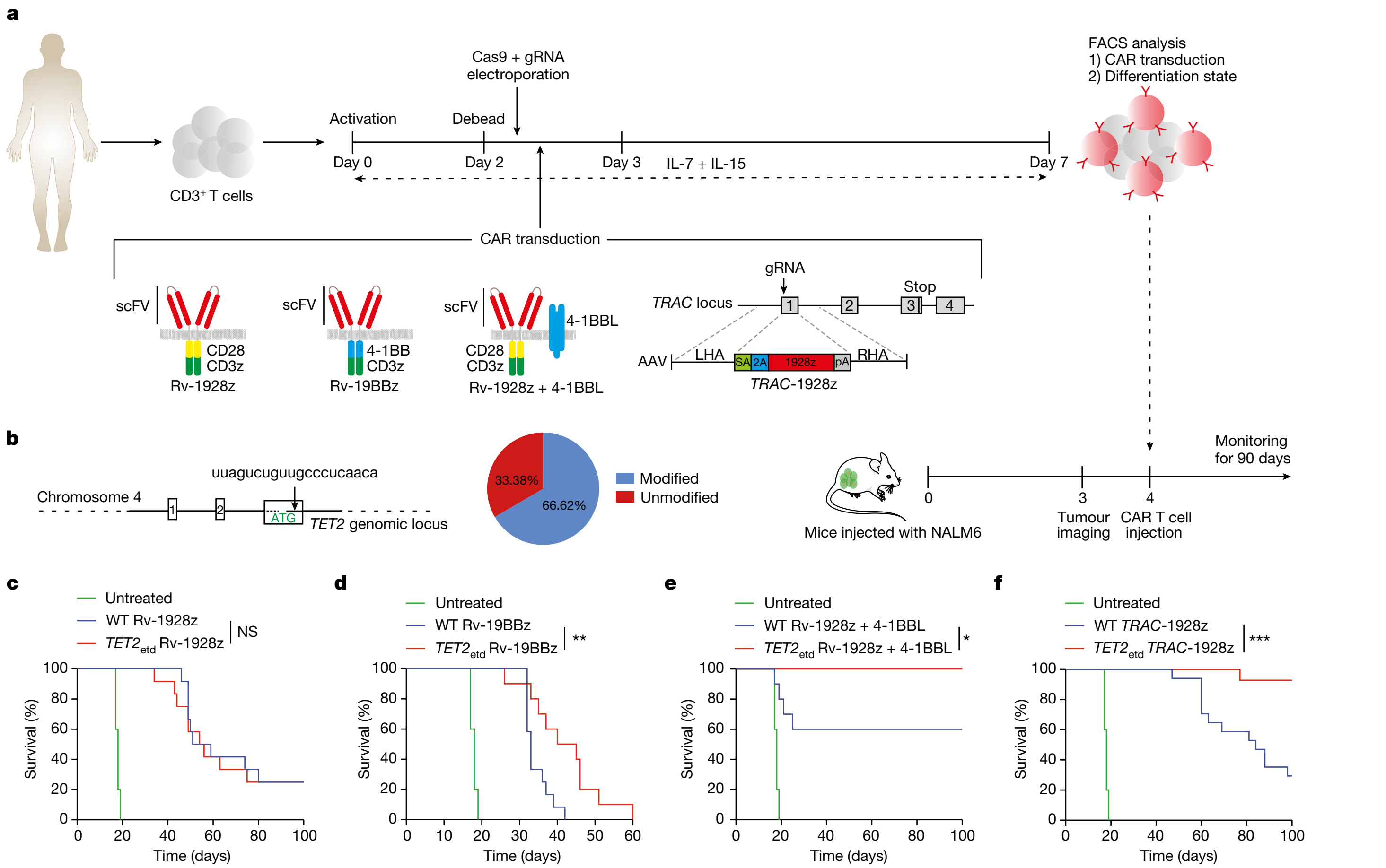


Fig. 1 | Effect of *TET2* disruption on CAR T cell therapeutic efficacy is dependent on CAR design. a,**b**, Schematics of in vitro CAR T cell generation and the NALM6 xenograft mouse model (**a**), and *TET2* targeting gRNA and editing efficiency (**b**). **c**,**d**, Mice survival under Rv-1928z (dose: 1×10^5; $n=12$) (**c**) and Rv-19BBz (dose: 2×10^5; $n=12$) (**d**) CAR T cell treatments. **e**,**f**, Cancer-free survival of mice treated with Rv-1928z + 4-1BBL (dose: 5×10^4; $n=10$) (**e**) and *TRAC*-1928z (dose: 1×10^5; $n=15$) (**f**) CAR T cells. Data were collated from two donors. Untreated $n=5$. Log-rank Mantel–Cox test was used; $P<0.05$ was considered statistically significant. *P* values are denoted: not significant (NS) $P>0.05$, $^{*}P<0.05$, $^{**}P<0.01$ and $^{***}P<0.001$ (**c**–**f**). The human, mouse and lipid bilayer illustrations in part **a** were generated using Servier Medical Art, CC BY 3.0.

and TIM3) was indistinguishable between unedited and *TET2*-edited groups for both CAR designs (Extended Data Fig. 1i). Intrigued by the different outcome between the two CARs, we further evaluated the 1928z CAR design in two distinct contexts that extend its persistence, by either co-expressing 4-1BBL (Rv-1928z + 4-1BBL)[23] or by transcribing the CAR from the *TRAC* locus (*TRAC*-1928z)[24] (Fig. 1a). As with Rv-1928z and Rv-19BBz, *TET2* editing did not affect CAR transduction efficiency or the pre-infusion T cell phenotype of either CAR T cell populations (Extended Data Fig. 2a,b), but their efficacy was increased relative to their non-edited counterparts (Fig. 1e,f and Extended Data Fig. 2c,d). This increased efficacy was associated with increased expression of CCR7 in Rv-1928z + 4-1BBL and *TRAC*-1928z CAR T cells (Extended Data Fig. 2e,f). Although, it did not reach statistical significance for Rv-1928z + 4-1BBL (Extended Data Fig. 2f). Inhibitory receptor expression was similar between wild-type (WT) and *TET2*-edited groups for both Rv-1928z + 4-1BBL and *TRAC*-1928z CARs (Extended Data Fig. 2g). These findings thus established that disruption of *TET2* could augment therapeutic efficacy of either CAR, albeit depending on CAR expression.

Hyperproliferative CAR T cells emerge

Continued follow-up of these mice uncovered signs of clinical distress developing after 50 days in the absence of detectable tumour in mice treated with *TET2*-edited T cells (Fig. 2a and Extended Data Fig. 2c,d). Gross pathology revealed an enlarged spleen and liver, pale kidneys, and lungs with extensive T cell infiltration and absence of CD19^{+} leukaemia. The infiltrating T cells were CAR^{+} and Ki67^{+} (Fig. 2b). This prompted us to treat additional cohorts of mice with all four CAR T cell types (Rv-19BBz, Rv-1928z, Rv-1928z + 4-1BBL and *TRAC*-1928z), administering 2–5×10^5 CAR T cells to ensure tumour elimination in most mice to allow for long-term follow-up of all four groups (Fig. 2c). All CARs maintained long-term tumour remission as assessed by bioluminescence imaging (BLI), tumour was eliminated within 2–3 weeks of CAR T cell administration. Mice treated with Rv-CARs were euthanized on day 90 and *TRAC*-CAR T cells recipients on day 75. Bone marrow and splenic CAR T cell numbers were considerably increased in mice treated with *TET2*-edited Rv-19BBz, Rv-1928z + 4-1BBL and *TRAC*-1928z CAR T cells, compared with their unedited counterparts (Fig. 2d). CAR T cell numbers in recipients of *TET2*-edited and unedited Rv-1928z in both the bone marrow and the spleen, however, did not significantly differ (Fig. 2d), except for a single mouse (1 out of 10) that showed an increase in *TET2*-edited CAR T cells. Flow cytometric analysis of CAR T cells isolated from the bone marrow confirmed increased expression of CCR7 in *TET2*-edited Rv-19BBz, Rv-1928z + 4-1BBL and *TRAC*-1928z CAR T cells, but not in *TET2*-edited Rv-1928z CAR T cells (Extended Data Fig. 3a,b). Inhibitory receptor expression was again unchanged upon *TET2* editing across all four CAR designs (Extended Data Fig. 3c). This long-term follow-up thus confirmed that *TET2* editing increases therapeutic efficacy and T cell accumulation, but with pathological consequences appearing weeks or months after tumour clearance.

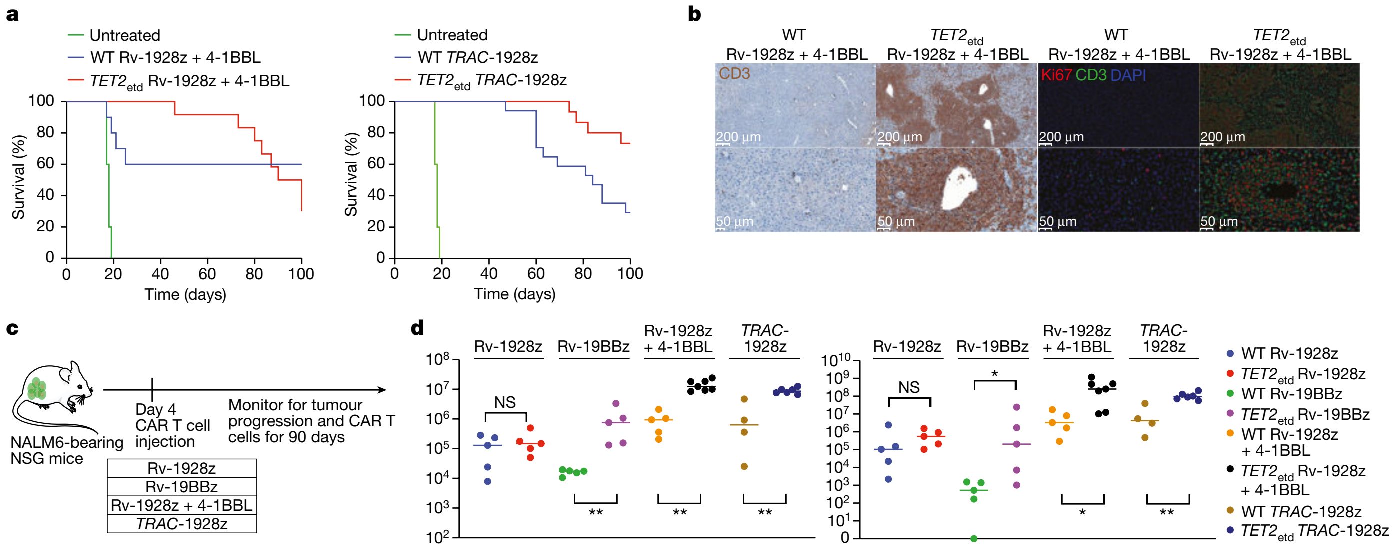


Fig. 2 | Effect of CAR design on long-term T cell accumulation upon CRISPR–Cas9 editing of the *TET2* locus. a, Overall survival of NALM6-bearing mice treated with Rv-1928z + 4-1BBL ($n = 10$) and *TRAC*-1928z ($n = 15$). **b**, Immunohistochemistry and immunofluorescence staining of a liver section of mice treated with WT Rv-1928z + 4-1BBL and $TET2_{etd}$ Rv-1928z + 4-1BBL at day 90. Across four CARs, immunohistochemistry and immunofluorescence were performed for 15 mice treated with WT CAR T cells and 20 mice treated with $TET2_{etd}$ CAR T cells. **c**, Schematics of long-term CAR T cell and tumour monitoring. Rv-1928z CAR T cells were used at a dose of 4×10^5, Rv-19BBz CAR T cells at a dose of 5×10^5, Rv-1928z + 4-1BBL CAR T cells at a dose of 2×10^5 and *TRAC*-1928z CAR T cells at a dose of 4×10^5. **d**, CAR T cell quantification in the bone marrow (left panel) and spleen (right panel). Bars show median values. Two-sided Mann–Whitney test was used ($n = 5$ (Rv-1928z), $n = 5$ (Rv-19BBz), $n = 5$ (WT Rv-1928z + 4-1BBL), $n = 7$ ($TET2_{etd}$ Rv-1928z + 4-1BBL), $n = 4$ (WT *TRAC*-1928z) and $n = 5$ ($TET2_{etd}$ *TRAC*-1928z)). $P < 0.05$ was considered statistically significant. P values are denoted: NS $P > 0.05$, $^*P < 0.05$ and $^{**}P < 0.01$. Exact P values are available in Supplementary Table 4. The mouse illustration in part **c** was generated using Servier Medical Art, CC BY 3.0.

To ascertain that acquisition of this hyperproliferative phenotype was not specific to a single tumour model or a particular guide RNA (gRNA), we established a human prostate cancer model in NSG mice (Extended Data Fig. 4a), targeting prostate-specific membrane antigen (PSMA) in PC3-bearing mice with PSMA-28z + 4-1BBL CAR T cells. Peripheral blood PSMA-28z + 4-1BBL CAR T cells edited with either gRNA-g1 or gRNA-g2 were tenfold more abundant than control PSMA-28z + 4-1BBL CAR T cells edited with a scrambled gRNA by day 30 (Extended Data Fig. 4b). Splenic CAR T cell quantification revealed over 100 million CAR T cells per spleen in recipients of *TET2*-edited PSMA-28z + 4-1BBL by day 45 (Extended Data Fig. 4c), establishing that late acquisition of a hyperproliferative phenotype is not specific to a tumour model or gRNA.

As Cas9-mediated *TET2* editing resulted in either unedited, monoallelic or biallelic disruption in individual T cells, we could test whether total loss of TET2 is required for achieving sustained proliferation. For this analysis, we focused on two Rv-1928z CAR populations yielding different rates of T cell accumulation: Rv-1928z and Rv-1928z + 4-1BBL (Extended Data Figs. 5a and 7a). Pre-infusion, Rv-1928z and Rv-1928z + 4-1BBL CAR T cells showed similar *TET2*-editing efficiency (Fig. 3a,b). By day 21 post-infusion, *TET2* editing was enriched for Rv-1928z + 4-1BBL but not Rv-1928z (Extended Data Fig. 5b). In subsequent follow-up, very high T cell counts were reached in 12 out of 15 mice treated with *TET2*-edited Rv-1928z + 4-1BBL, but only in 2 of 15 mice treated with *TET2*-edited Rv-1928z CAR T cells, becoming apparent by day 90 and day 200. In these latter two cases (2-2 and 2-00), we found a 19-bp deletion in both alleles in 2-2 (Fig. 3e) and a biallelic integration of a partial retroviral vector fragment in 2-00 (Extended Data Fig. 5c). Five of the expanded *TET2*-edited Rv-1928z + 4-1BBL populations harvested at day 90 were randomly selected for analysis and all were found to be nearly entirely (more than 98%) biallelically *TET2*-edited (Fig. 3f and Extended Data Fig. 5d–g). Western blot analysis showed an absence of TET2 protein in biallelically edited CAR T cells (Extended Data Fig. 5h). Thus, biallelic *TET2* editing ($TET2_{bed}$) is enriched over time, irrespective of CAR design, consistent with it being required for achieving a hyperproliferative T cell state.

We assessed clonal composition in the hyperproliferative CAR T cell populations by TCRvβ sequencing. All five Rv-1928z + 4-1BBL populations were multiclonal, with no clone constituting more than 50% of the total CAR product, except for sample 17-1 in which a single clone accounted for approximately 82% of the CAR T cells (Fig. 3f and Extended Data Fig. 5d–g). By contrast, both Rv-1928z populations (2-2 and 2-00) largely consisted in a single clone (more than 95%; Fig. 3e and Extended Data Fig. 5c), consistent with the lesser probability of 1928z CAR T cells achieving clonal expansion. TCRvβ sequencing of hyperproliferative *TRAC*-1928z and Rv-19BBz also revealed multiclonal expansion (Extended Data Fig. 5i,j).

Lack of shared TCRs between different hyperproliferative populations, the absence of graft-versus-host disease in mice bearing hyperproliferative CAR T cell population and the emergence of clonal dominance in *TRAC*-1928z CAR T cell-treated mice (in which CAR T cells lack TCR expression[24]) strongly suggested that the TCR is not required for acquisition of a hyperproliferative phenotype. To further exclude a role for TCR in sustained clonal expansion, we ablated TCR expression in conjunction with *TET2* disruption before transduction of Rv-1928z + 4-1BBL and compared the frequency of emergence of the hyperproliferative phenotype in recipient mice. Long-term follow-up of TCR^+ *TET2*-edited and TCR^- *TET2*-edited Rv-1928z + 4-1BBL CAR T cells revealed no differences in frequency of CAR T cells achieving a hyperproliferative state and their differentiation state (Extended Data Fig. 6a–c), confirming that TCR is not required for sustained proliferation.

We hypothesized that the increased expansion and clonal diversity imparted by Rv-19BBz, Rv-1928z + 4-1BBL and *TRAC*-1928z, contrasting with that of Rv-1928z, were owed to the CAR and not only *TET2* editing (Fig. 3e,f and Extended Data Fig. 5d–j). To this end, we introduced either Rv-1928z or Rv-1928z + 4-1BBL in the same pool of *TET2*-edited T cells and compared the fate of common TCRvβ clonotypes expressing either CAR (Extended Data Fig. 7a). Pairwise analysis of the same TCRvβ clonotypes in different mice revealed major differences in clonal evolution between Rv-1928z and Rv-1928z + 4-1BBL from day 0

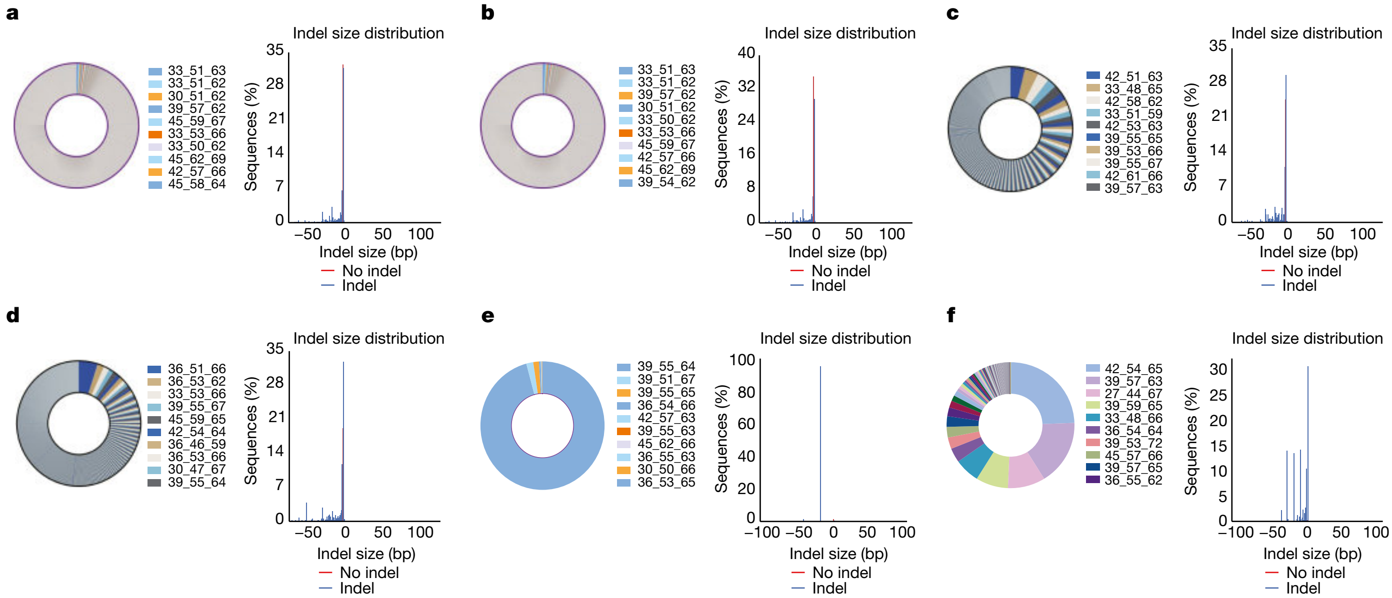


Fig. 3 | Hyperproliferative *TET2*-edited CAR T populations are oligoclonal and biallelically edited for *TET2*. a,**b**, Pre-infusion TCRvβ sequencing (left panel) and *TET2* status (right panel) of Rv-1928z (**a**) and Rv-1928z + 4-1BBL (**b**). CAR T cells were generated from the same donor. Indels include insertions (+) and deletions (−). **c**,**d**, TCRvβ sequencing (left panel) and *TET2* status (right panel) of Rv-1928z (**c**) and Rv-1928z + 4-1BBL (**d**). CAR T cells were isolated at week 3 post-infusion in mice. **e**,**f**, TCRvβ sequencing (left panel) and *TET2* status (right panel) of hyperproliferative Rv-1928z (2-2; **e**) and Rv-1928z + 4-1BBL (15-1; **f**). The CAR T cell population reveals that multiple clones from the pre-infusion population can become hyperproliferative but they require biallelic *TET2* editing. Hyperproliferative CAR T cells were isolated at day 90.

(pre-infusion CAR T cells) to day 21 (Extended Data Fig. 7b,c). This divergent evolution is illustrated by tracking the persistence of the 100 most frequent clones in the Rv-1928z pre-infusion cell population, all of which were also present in the Rv-1928z + 4-1BBL pre-infusion product (Extended Data Fig. 7d). By day 21, most (70 out of 100) of these clones were still detected in Rv-1928z + 4-1BBL CAR T cells, whereas only 3 out of 100 were detectable in recipients of Rv-1928z CAR T cells (Extended Data Fig. 7d). By retro-tracking clones present in hyperproliferative populations (day 90) to pre-infusion, we found few persisting clones for Rv-1928z in contrast to Rv-1928z + 4-1BBL (Extended Data Fig. 7e,f), even though both Rv-1928z and Rv-1928z + 4-1BBL had similar pre-infusion clonal diversity (Extended Data Fig. 7g). The difference between Rv-1928z and Rv-1928z + 4-1BBL in their respective clonal longevity was further evidenced by tracking the 100 most frequent shared clones from the pre-infusion Rv-1928z and Rv-1928z + 4-1BBL CAR populations up to day 90. None was detected in Rv-1928z (Extended Data Fig. 7h), whereas some of the earliest clones detected on day 0 in the Rv-1928z + 4-1BBL population remained detectable by day 90 (Extended Data Fig. 7i), although they were not dominant (Extended Data Fig. 7j). These tracking data confirmed that the probability of a given clonotype acquiring a hyperproliferative phenotype upon loss of TET2 is determined by the CAR and that the relative resistance imparted by Rv-1928z could be overcome on engaging the 4-1BB pathway by overexpressing 4-1BBL.

Reduced effector function in CAR T cells

To assess the functional properties of the hyperproliferative CAR T cell population, we first evaluated the cytolytic function of hyperproliferative $TET2_{bed}$ CAR T cells in vitro and in vivo. $TET2_{bed}$ CAR T cells demonstrated diminished cytolytic ability and were relatively ineffective for eliminating established NALM6 in vivo (Fig. 4a and Extended Data Fig. 8a,b), requiring a higher CAR T cell dosage to delay tumour progression (Extended Data Fig. 8c). $TET2_{bed}$ CAR T cells showed a profound loss of effector cytokine secretion upon activation (Fig. 4b). For further molecular characterization, we focused on Rv-1928z + 4-1BBL CAR T cells as the unedited Rv-1928z + 4-1BBL CAR T cells persisted the most among the four tested CAR designs and thus could provide a matched, unedited control. Transcriptional profiling of hyperproliferative $TET2_{bed}$ and WT Rv-1928z + 4-1BBL CAR T cells revealed an increased expression of cell-cycle-related factors in the former (Fig. 4c,d). $TET2_{bed}$ Rv-1928z + 4-1BBL CAR T cells demonstrated diminished effector cytokine induction upon activation (Fig. 4e and Extended Data Fig. 8d), which led us to further examine effector function in WT and $TET2_{etd}$ CAR T cells over multiple rounds of antigen stimulation (Extended Data Fig. 8e). Early on, $TET2_{etd}$ and WT CAR T cells displayed indistinguishable cytolytic capacity and effector cytokine secretion (Extended Data Fig. 8f–i). However, after multiple rounds of stimulation (five stimulations over 14 days), $TET2_{etd}$ CAR T cells exhibited reduced cytolytic function and effector cytokine secretion compared with WT CAR T cells (Extended Data Fig. 8j,k). Collectively, these observations establish that TET2 deficiency leads to a gradual erosion of effector function but predisposes to the emergence of $TET2_{bed}$ CAR T cell clones that are characterized by sustained proliferation, moderate cytolytic potential and poor cytokine responses.

Consistent with this functional profile, we did not find expression of the memory-associated transcription factor TCF1 (Extended Data Fig. 8l) or an enrichment of memory gene sets in $TET2_{bed}$ compared with WT Rv-1928z + 4-1BBL (Fig. 4f), despite the increased expression of some memory-associated biomarkers such as CCR7. Instead, we found enrichment in angioimmunoblastic T cell lymphoma and HTLV1-driven adult T cell leukaemia/lymphoma datasets (Fig. 4g). This led us to search for potential genetic drivers of proliferation and investigate the proliferative potential of $TET2_{bed}$ CAR T cells upon secondary transplant.

BATF3 drives hyperproliferation

To assess whether $TET2_{bed}$ clones had acquired mutations that could account for their clonal dominance, we performed whole-exome sequencing in three clones expressing different CARs (Extended Data Fig. 9a,c,e). Numerous non-synonymous point mutations were observed in all three dominant clones (Extended Data Fig. 9b,d,f).

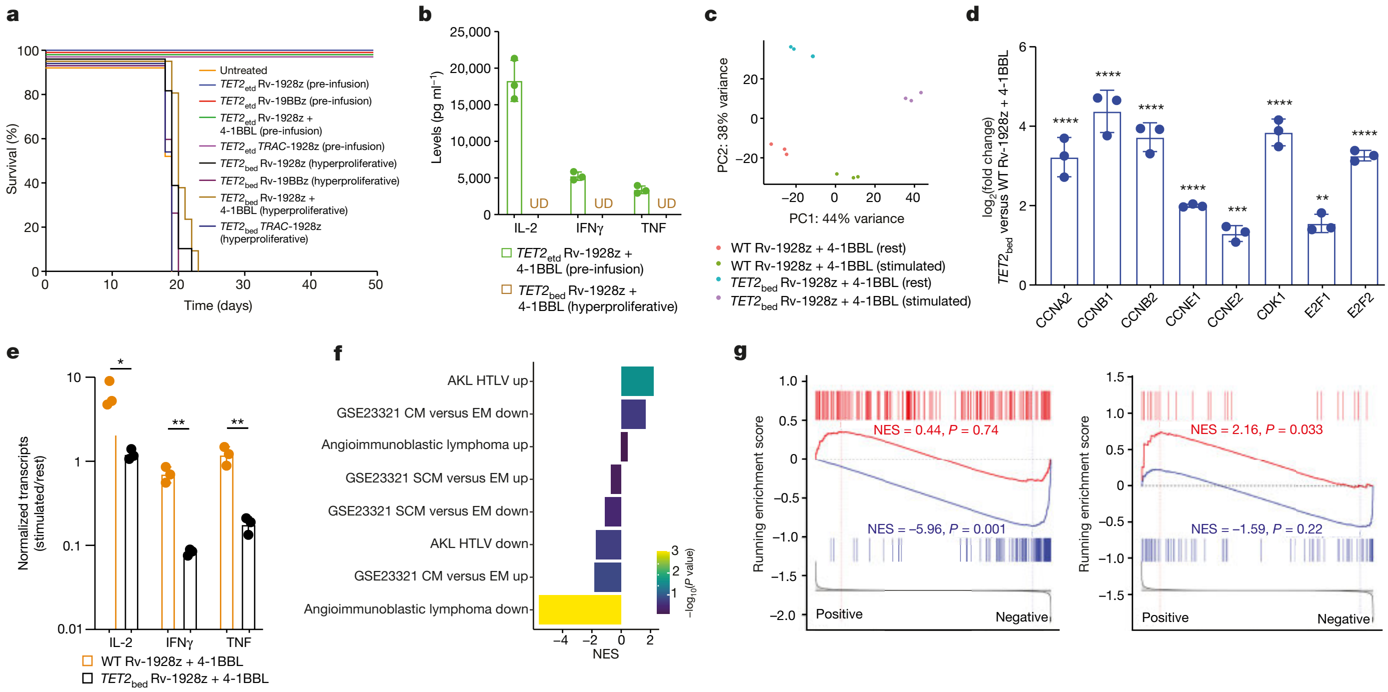


Fig. 4 | Loss of effector function in hyperproliferative *TET2*$_{bed}$ CAR T cells. **a**, NALM6-bearing NSG mice were either treated with 5×10^5 hyperproliferative *TET2*$_{bed}$ Rv-1928z ($n = 7$), Rv-19BBz ($n = 3$), Rv-1928z + 4-1BBL ($n = 7$) and *TRAC*-1928z ($n = 5$) CAR T cells or pre-infusion *TET2*$_{etd}$ Rv-1928z (dose: 4×10^5), Rv-19BBz (dose: 5×10^5), Rv-1928z + 4-1BBL (dose: 2×10^5) and *TRAC*-1928z (dose: 4×10^5) ($n = 5$ for all pre-infusion *TET2*-edited CAR T cells). **b**, Effector cytokine secretion upon activation of pre-infusion *TET2*$_{etd}$ and hyperproliferative Rv-1928z + 4-1BBL population. Data are represented as mean ± s.d. ($n = 3$). **c**, Principal component (PC) analysis of resting and stimulated (24-h after co-culture with CD3/CD28 beads at 1:1 bead-to-cell ratio) WT Rv-1928z + 4-1BBL and *TET2*$_{bed}$ Rv-1928z + 4-1BBL. **d**, Elevated levels of cell cycle factors in *TET2*$_{bed}$ Rv-1928z + 4-1BBL compared with WT Rv-1928z + 4-1BBL. Data are represented as mean ± s.d. ($n = 3$). *P* values were determined by Wald test with false discovery rate correction (two-sided). **e**, Reduced induction of effector cytokines in response to CD3/CD28 bead stimulation in *TET2*$_{bed}$ Rv-1928z + 4-1BBL compared with WT Rv-1928z + 4-1BBL. Data are represented as mean ± s.d. ($n = 3$). *P* values were determined by two-sided Student's *t*-test with false discovery rate correction. **f**,**g**, Gene set enrichment analysis reveals no enrichment in central memory (CM) and stem cell memory (SCM) compartments for *TET2*$_{bed}$ Rv-1928z + 4-1BBL compared with WT Rv-1928z + 4-1BBL (**f**), and enrichment in angioimmunoblastic T cell lymphoma (left panel) and HTLV1-driven adult T cell leukaemia/lymphoma (right panel) gene sets of *TET2*$_{bed}$ Rv-1928z + 4-1BBL (**g**). Red line indicates enrichment profile in the upregulated gene dataset and blue line indicates enrichment profile in the downregulated gene dataset in **g**. *P* values were corrected for multiple comparisons by the BKY method. $P < 0.05$ was considered statistically significant. *P* values are denoted: NS $P > 0.05$, *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$ and ****$P < 0.0001$. Exact *P* values are available in Supplementary Table 4. UD, undetected; EM, effector memory; NES, normalized enrichment score.

Analysis of translocations for these three samples only identified CAR (CD28–CD3z) fusions (Supplementary Table 1). Some chromosomal amplifications and megabase-scale deletions were observed in a subset of the dominant clone population in samples 17-1 and 4-1 (Extended Data Fig. 9a,e). Given their substantially lower frequency than that of the dominant clone, these gross chromosomal defects appeared to be late occurring secondary events. For the retroviral-encoded CARs in samples 17-1 and 2-2, we identified the sites of retroviral integration. None of them disrupted or integrated next to cancer-related genes associated with angioimmunoblastic T cell lymphoma or T cell lymphoma (Supplementary Table 2). Together, we found that hyperproliferative *TET2*$_{bed}$ T cells are prone to acquiring somatic mutations, but do not bear recurrent genetic mutations or mutations known to be associated with T cell malignancies.

Secondary transplant studies of *TET2*$_{bed}$ CAR T cells have shown that they did not engraft on their own, but could persist with exogenous cytokine supplementation, promptly declining after cessation of cytokine administration (Extended Data Fig. 10a,b). Cell numbers remained modest and were barely detectable at day 150 when the study reached its intended end point (Extended Data Fig. 10c). These findings indicate that *TET2*$_{bed}$ CAR T cells are unable to autonomously sustain their proliferation upon secondary transplant.

The lack of a conserved genetic driver of proliferation of *TET2*$_{bed}$ CAR T cells prompted us to study whether their epigenetic state enables sustained proliferation. Assay for transposase-accessible chromatin using sequencing analysis revealed significant differences between accessible chromatin regions of WT and *TET2*$_{bed}$ Rv-1928z + 4-1BBL CAR T cells (Supplementary Fig. 1a). The AP-1 family binding motif was the most significantly enriched motif in differentially open chromatin regions of *TET2*$_{bed}$ CAR T cells (Fig. 5a). Transcriptional analyses in these same cells revealed that, among the AP-1 factors, *BATF3* was the most significantly upregulated in *TET2*$_{bed}$ CAR T cells (Fig. 5b). BATF3 has been previously implicated as a driver of proliferation in T cell leukaemia/lymphoma[25–27] in part by inducing a *MYC* transcriptional program[25,26]. Distinct promoter and gene body regions of *BATF3*, with some encompassing consensus AP-1-binding motifs, were found to be more readily accessible in hyperproliferative *TET2*$_{bed}$ Rv-1928z + 4-1BBL CAR T cells than WT Rv-1928z + 4-1BBL CAR T cells (Fig. 5c,d and Supplementary Fig. 1b). *TET2*$_{bed}$ Rv-1928z + 4-1BBL CAR T cells showed a strong enrichment in hallmark MYC targets when compared with WT Rv-1928z + 4-1BBL CAR T cells (Fig. 5e). Flow cytometric analyses of unedited and hyperproliferative CAR T cells isolated at day 90 showed a higher fraction of BATF3$^+$MYC$^+$ in hyperproliferative CAR T cells (Fig. 5f,g). Analysis of *BATF3* and *MYC* expression upon *TET2* editing in CAR T cells (Supplementary Fig. 1c) revealed that CAR activation induced *BATF3* expression (Supplementary Fig. 1d). The levels of BATF3 and MYC did not differ between WT and *TET2*-edited CAR T cells at early time points (days 1 and 8) (Supplementary Fig. 1e), but increased after five rounds of

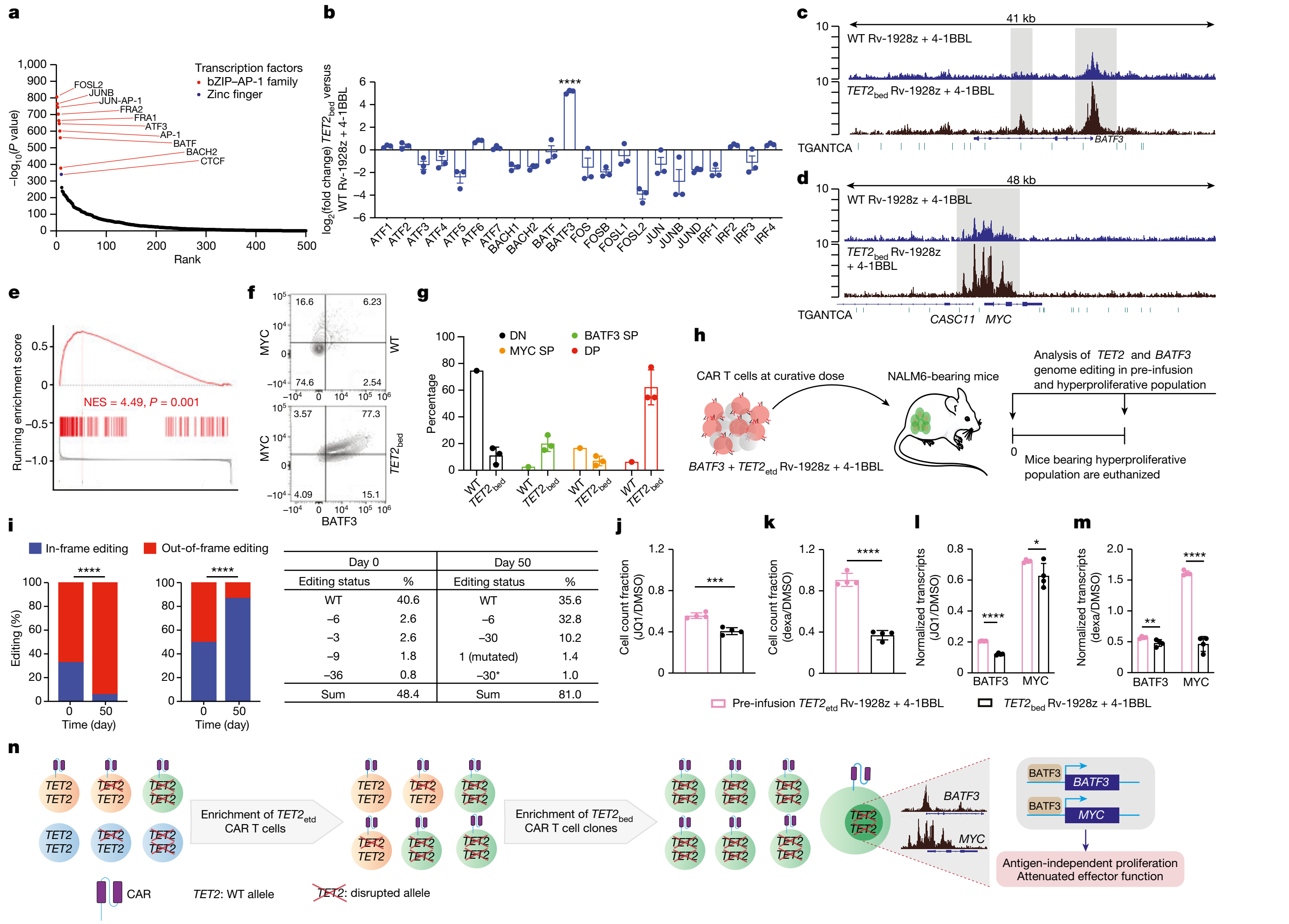


Fig. 5 | The BATF3–MYC axis drives hyperproliferation of *TET2*$_{bed}$ CAR T cells. a, The AP-1-binding motif was most significantly enriched in the open chromatin region of *TET2*$_{bed}$ Rv-1928z + 4-1BBL. **b**, RNA expression of the AP-1 family of transcription factors in *TET2*$_{bed}$ Rv-1928z + 4-1BBL and WT Rv-1928z + 4-1BBL. Data are represented as mean ± s.d. ($n = 3$). *P* values were determined by Wald test with false discovery rate correction (two-sided). **c**,**d**, Increased genomic accessibility (highlighted by the grey background) in promoter and gene body regions of *BATF3* (**c**) and *MYC* (**d**). The AP-1-binding motif is marked by green dashes. **e**, Gene set enrichment analysis reveals increased *MYC* signalling (Hallmark_MYC_V1, M5926) in *TET2*$_{bed}$ Rv-1928z + 4-1BBL compared with WT Rv-1928z + 4-1BBL. *P* values were corrected for multiple comparisons by the BKY method. **f**,**g**, Flow cytometry for BATF3 and MYC in WT and *TET2*$_{bed}$ Rv-1928z + 4-1BBL CAR T cells at day 90 (**f**). The WT sample was pooled from ten mice. The *TET2*$_{bed}$ sample is a representative population from one mouse in **f**. Data in **g** are a summary from three mice and are presented as mean ± s.d. **h**, Schematics of the *TET2* and *BATF3* dual-editing study. **i**, *TET2*-editing (left panel) and *BATF3*-editing (right panel and table) outcomes were determined at pre-infusion and hyperproliferation. *P* values were determined by two-sided χ^2 test. **j**,**k**, Cells were either treated with DMSO, JQ1 (500 nM) or dexamethasone (dexa; 1 µm). DMSO-normalized cell counts for JQ1 (**j**) and dexa (**k**) are shown. Data are represented as mean ± s.d. ($n = 4$). *P* values were determined by two-sided, unpaired *t*-test. **l**,**m**, Quantitative PCR study for *BATF3* and *MYC* under JQ1 and dexa treatment. Transcripts were normalized to *B2M*. DMSO-normalized *BATF3* and *MYC* levels under JQ1 treatment (**l**) and dexa (**m**) are shown. Data are represented as mean ± s.d. ($n = 4$). *P* values were determined by two-sided, multiple unpaired *t*-tests corrected by the BKY method. **n**, Graphical model summarizing the results. SP, single positive; DN, double negative; DP, double positive. $P < 0.05$ was considered statistically significant. *P* values are denoted: NS $P > 0.05$, *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$ and ****$P < 0.0001$. Exact *P* values are available in Supplementary Table 4. The mouse illustration in part **h** was generated using Servier Medical Art, CC BY 3.0.

stimulation (day 15) in the *TET2*-edited group (BATF3, Supplementary Fig. 1f; MYC, Supplementary Fig. 1g). These observations suggest that TET2 deficiency gradually establishes an epigenetic state conducive to increased *BATF3* and *MYC* expression that may ultimately result in the sustained proliferation of *TET2*$_{bed}$ CAR T cell clones.

To directly test the role of *BATF3* in acquisition of the hyperproliferative state, we designed an in vivo study in which *BATF3* and *TET2* are both edited in Rv-1928z + 4-1BBL CAR T cells (Supplementary Fig. 2a), hypothesizing that in-frame *BATF3* edits would be enriched, and out-of-frame edits would be counter-selected over time. NALM6-bearing mice were treated with a predictably curative dose of *BATF3* and *TET2*$_{etd}$ Rv-1928z + 4-1BBL CAR T cells to allow for long-term monitoring (Fig. 5h). Deep sequencing at both *TET2* and *BATF3* loci indeed confirmed an enrichment of out-of-frame edits at the *TET2* locus and in-frame edits at the *BATF3* locus when the hyperproliferative population emerged at day 50 (Fig. 5i), confirming the essential requirement for *BATF3* expression to acquire the hyperproliferative phenotype.

We further corroborated this dependency on BATF3 pharmacologically. JQ1 is an inhibitor of the BET protein BRD4, which has been previously shown to inhibit *BATF3* and *MYC* expression in adult T cell leukaemia/lymphoma cells[26]. Although JQ1 inhibited proliferation of

all tested CAR populations, *TET2*$_{bed}$ CAR T cells were more sensitive to JQ1 treatment than were pre-infusion *TET2*-edited CAR T cells (Fig. 5j and Supplementary Fig. 2b,c). This heightened sensitivity to JQ1 was associated with a greater suppression of *BATF3* and *MYC* expression in *TET2*$_{bed}$ CAR T cells (Fig. 5l and Supplementary Fig. 2d,e). Dexamethasone has been shown to suppress AP-1 factors[28,29]. In contrast to JQ1, dexamethasone did not limit proliferation of pre-infusion *TET2*-edited CAR T cells (Fig. 5k and Supplementary Fig. 2e). However, it markedly inhibited proliferation of *TET2*$_{bed}$ CAR T cells (Fig. 5k and Supplementary Fig. 2e). This increased sensitivity to dexamethasone was associated with reduction in expression of both *BATF3* and *MYC* in *TET2*$_{bed}$ CAR T cells (Fig. 5m and Supplementary Fig. 2f). By contrast, *MYC* expression remained elevated in pre-infusion *TET2*-edited CAR T cells despite *BATF3* inhibition (Fig. 5m and Supplementary Fig. 2f). This differential expression of *MYC* between pre-infusion *TET2*-edited CAR T cells and *TET2*$_{bed}$ CAR T cells upon exposure to dexamethasone further supports the dependency of *MYC* expression on BATF3 in *TET2*$_{bed}$ CAR T cells.

Discussion

We found that biallelic *TET2* disruption can enhance the efficacy and proliferation of human CAR T cells. The increased efficacy is consistent with earlier observations in mouse TCR transgenic T cells[9], and in a case report of a clonal expansion in a patient treated with a 4-1BB CAR[10]. The functional enhancement of CAR T cell antitumour activity, however, depends on the CAR and not merely the *TET2* status. Tumour elimination by *TET2*-edited T cells was enhanced with Rv-19BBz, Rv-1928z + 4-1BBL and *TRAC*-1928z, but not Rv-1928z. Over time, *TET2*$_{bed}$ CAR T cells repeatedly emerged in a hyperproliferative state, consistently associated with CAR expression, albeit not requiring the presence of CAR antigen, and independently of the TCR. TCRvβ sequencing identified multiple T cell clones in mice treated with *TET2*-edited Rv-19BBz, Rv-1928z + 4-1BBL and *TRAC*-1928z, whereas hyperproliferative *TET2*-edited Rv-1928z CAR T cells were rare and monoclonal when they occurred. These observations underscore a probabilistic fate, in which the chance of establishing a hyperproliferative state in *TET2*-edited T cells is low in shorter-lived Rv-1928z CAR T cells, rarely allowing rare breakthrough clones, and increased with CAR designs that autonomously promote greater persistence. Hyperproliferative *TET2*$_{bed}$ CAR T cells bear secondary mutations, consistent with previous reports that have shown a role for TET2 in maintaining genomic integrity[30,31]. However, we did not identify a recurrent mutation among different *TET2*$_{bed}$ CAR T cell populations or mutations known to be associated with T cell leukaemia/lymphoma[22]. Instead, we identified a strict requirement for BATF3 expression, associated with an epigenetic signature characterized by enhanced *BATF3* and *MYC* accessibility.

AP-1 factors are critically involved in distinct T cell states[32–35]. *Batf* overexpression in mouse CAR T cells enhances their antitumour activity[33]. *BATF3* overexpression in T cells enhances their CCR7 expression and memory formation[36,37], although high levels of BATF3 has also been associated with human CAR T cell exhaustion[32]. We found here that in the context of epigenetic changes brought on by loss of TET2, sustained *BATF3* expression programs a hyperproliferative state rather than T cell memory. Furthermore, *TET2*$_{bed}$ CAR T cells demonstrate reduced cytolytic function and poor cytokine response upon activation, despite maintaining genome accessibility in effector loci (Supplementary Fig. 3), which suggests that effector functions are transcriptionally downregulated. These observations point to a T cell state that differs from both canonically defined T cell exhaustion[15,16] and T cell memory[13,14]. Our findings thus establish TET2 as an epigenetic regulator of *BATF3* to prevent unchecked proliferation and maintain T cell genomic integrity.

Several AP-1 factors are known to potentially promote oncogenesis[38]. *JUN* and *BATF* overexpression can lead to uncontrolled proliferation[39,40]. BATF3 has been shown to drive proliferation in T cell leukaemia/lymphoma through MYC[26] or IL-2R[27]. Sustained *BATF3* expression in T cell leukaemia/lymphoma is associated with activated super-enhancers at the *BATF3* locus[26,27]. The hyperproliferative CAR T cell phenotype that we report here underscores the potency of CAR T cell epigenetic programming but reveals long-term safety concerns that may arise from manipulating *TET2* (ref. [41]) and AP-1 factors[38]. Remarkably, however, *TET2*$_{bed}$ CAR T cells remained highly sensitive to dexamethasone, which lowered the expression of both *BATF3* and *MYC* in *TET2*$_{bed}$ CAR T cells. This high sensitivity may explain the sudden clonal contraction upon corticosteroid administration to manage cytokine release syndrome observed in the patient bearing a TET2-deficient 19BBz CAR T cell clone[10]. The intentional disruption of *TET2* for CAR T cell therapy may nonetheless be concerning, especially in elderly individuals who are more likely to have mutations in *DNMT3A*[42], which can synergize with loss of TET2 to precipitate T cell oncogenesis[43]. Screening for pre-existing mutations that predispose to hyperproliferation or transformation should help to mitigate this hazard. Transient or partial suppression of *TET2* during CAR T cell production[44] may eschew such a risk.

In summary, disruption of *TET2* enhances CAR T cell efficacy and promotes sustained T cell accumulation but exposes to the risk of a hyperproliferative state that is prone to accumulating secondary mutations. These findings demonstrate the formidable potential of epigenetic reprogramming to alter CAR T cell fate and highlight how an AP-1 factor, such as BATF3, may direct distinct effector and proliferative states under different epigenetic contexts.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-05692-z.

1. Kakarla, S. & Gottschalk, S. CAR T cells for solid tumors: armed and ready to go? *Cancer J.* **20**, 151–155 (2014).
2. Sadelain, M., Riviere, I. & Riddell, S. Therapeutic T cell engineering. *Nature* **545**, 423–431 (2017).
3. June, C. H. & Sadelain, M. Chimeric antigen receptor therapy. *N. Engl. J. Med.* **379**, 64–73 (2018).
4. Guedan, S., Calderon, H., Posey, A. D. Jr & Maus, M. V. Engineering and design of chimeric antigen receptors. *Mol. Ther. Methods Clin. Dev.* **12**, 145–156 (2019).
5. Globerson Levin, A., Riviere, I., Eshhar, Z. & Sadelain, M. CAR T cells: building on the CD19 paradigm. *Eur. J. Immunol.* **51**, 2151–2163 (2021).
6. Philip, M. et al. Chromatin states define tumour-specific T cell dysfunction and reprogramming. *Nature* **545**, 452–456 (2017).
7. Khan, O. et al. TOX transcriptionally and epigenetically programs $CD8^+$ T cell exhaustion. *Nature* **571**, 211–218 (2019).
8. Pastor, W. A., Aravind, L. & Rao, A. TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nat. Rev. Mol. Cell Biol.* **14**, 341–356 (2013).
9. Carty, S. A. et al. The loss of TET2 promotes $CD8^+$ T cell memory differentiation. *J. Immunol.* **200**, 82–91 (2018).
10. Fraietta, J. A. et al. Disruption of TET2 promotes the therapeutic efficacy of CD19-targeted T cells. *Nature* **558**, 307–312 (2018).
11. Bowman, R. L. & Levine, R. L. TET2 in normal and malignant hematopoiesis. *Cold Spring Harb. Perspect. Med.* **7**, a026518 (2017).
12. Chiba, S. Dysregulation of TET2 in hematologic malignancies. *Int. J. Hematol.* **105**, 17–22 (2017).
13. Kaech, S. M., Wherry, E. J. & Ahmed, R. Effector and memory T-cell differentiation: implications for vaccine development. *Nat. Rev. Immunol.* **2**, 251–262 (2002).
14. Farber, D. L., Yudanin, N. A. & Restifo, N. P. Human memory T cells: generation, compartmentalization and homeostasis. *Nat. Rev. Immunol.* **14**, 24–35 (2014).
15. Wherry, E. J. T cell exhaustion. *Nat. Immunol.* **12**, 492–499 (2011).
16. Blank, C. U. et al. Defining 'T cell exhaustion'. *Nat. Rev. Immunol.* **19**, 665–674 (2019).
17. Majzner, R. G. & Mackall, C. L. Clinical lessons learned from the first leg of the CAR T cell journey. *Nat. Med.* **25**, 1341–1355 (2019).
18. Rafiq, S., Hackett, C. S. & Brentjens, R. J. Engineering strategies to overcome the current roadblocks in CAR T cell therapy. *Nat. Rev. Clin. Oncol.* **17**, 147–167 (2020).
19. Tahiliani, M. et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930–935 (2009).
20. Delhommeau, F. et al. Mutation in TET2 in myeloid cancers. *N. Engl. J. Med.* **360**, 2289–2301 (2009).
21. Tefferi, A., Lim, K. H. & Levine, R. Mutation in TET2 in myeloid cancers. *N. Engl. J. Med.* **361**, 1117–1118 (2009).

22. Watatani, Y. et al. Molecular heterogeneity in peripheral T-cell lymphoma, not otherwise specified revealed by comprehensive genetic profiling. *Leukemia* **33**, 2867–2883 (2019).
23. Zhao, Z. et al. Structural design of engineered costimulation determines tumor rejection kinetics and persistence of CAR T cells. *Cancer Cell* **28**, 415–428 (2015).
24. Eyquem, J. et al. Targeting a CAR to the TRAC locus with CRISPR/Cas9 enhances tumour rejection. *Nature* **543**, 113–117 (2017).
25. Lollies, A. et al. An oncogenic axis of STAT-mediated BATF3 upregulation causing MYC activity in classical Hodgkin lymphoma and anaplastic large cell lymphoma. *Leukemia* **32**, 92–101 (2018).
26. Nakagawa, M. et al. Targeting the HTLV-I-regulated BATF3/IRF4 transcriptional network in adult T cell leukemia/lymphoma. *Cancer Cell* **34**, 286–297.e10 (2018).
27. Liang, H. C. et al. Super-enhancer-based identification of a BATF3/IL-2R-module reveals vulnerabilities in anaplastic large cell lymphoma. *Nat. Commun.* **12**, 5577 (2021).
28. Gonzalez, M. V. et al. Glucocorticoids antagonize AP-1 by inhibiting the activation/phosphorylation of JNK without affecting its subcellular distribution. *J. Cell Biol.* **150**, 1199–1208 (2000).
29. Patil, R. H. et al. Dexamethasone inhibits inflammatory response via down regulation of AP-1 transcription factor in human lung epithelial cells. *Gene* **645**, 85–94 (2018).
30. Kafer, G. R. et al. 5-Hydroxymethylcytosine marks sites of DNA damage and promotes genome stability. *Cell Rep.* **14**, 1283–1292 (2016).
31. Chen, L. L. et al. SNIP1 recruits TET2 to regulate c-MYC target genes and cellular DNA damage response. *Cell Rep.* **25**, 1485–1500.e4 (2018).
32. Lynn, R. C. et al. c-Jun overexpression in CAR T cells induces exhaustion resistance. *Nature* **576**, 293–300 (2019).
33. Seo, H. et al. BATF and IRF4 cooperate to counter exhaustion in tumor-infiltrating CAR T cells. *Nat. Immunol.* **22**, 983–995 (2021).
34. Man, K. et al. The transcription factor IRF4 is essential for TCR affinity-mediated metabolic programming and clonal expansion of T cells. *Nat. Immunol.* **14**, 1155–1165 (2013).
35. Man, K. et al. Transcription factor IRF4 promotes $CD8^+$ T cell exhaustion and limits the development of memory-like T cells during chronic infection. *Immunity* **47**, 1129–1141.e5 (2017).
36. McCutcheon, S. et al. CRISPR-based epigenome editing screens in primary human T cells. *Mol. Ther.* **30**, 165–166 (2022).
37. Ataide, M. A. et al. BATF3 programs $CD8^+$ T cell memory. *Nat. Immunol.* **21**, 1397–1407 (2020).
38. Eferl, R. & Wagner, E. F. AP-1: a double-edged sword in tumorigenesis. *Nat. Rev. Cancer* **3**, 859–868 (2003).
39. Schreiber, M. et al. Control of cell cycle progression by c-Jun is p53 dependent. *Genes Dev.* **13**, 607–619 (1999).
40. Logan, M. R., Jordan-Williams, K. L., Poston, S., Liao, J. & Taparowsky, E. J. Overexpression of Batf induces an apoptotic defect and an associated lymphoproliferative disorder in mice. *Cell Death Dis.* **3**, e310 (2012).
41. Quivoron, C. et al. TET2 inactivation results in pleiotropic hematopoietic abnormalities in mouse and is a recurrent event during human lymphomagenesis. *Cancer Cell* **20**, 25–38 (2011).
42. Couronne, L., Bastard, C. & Bernard, O. A. TET2 and DNMT3A mutations in human T-cell lymphoma. *N. Engl. J. Med.* **366**, 95–96 (2012).
43. Zhang, X. et al. DNMT3A and TET2 compete and cooperate to repress lineage-specific transcription factors in hematopoietic stem cells. *Nat. Genet.* **48**, 1014–1023 (2016).
44. Kong, W. et al. BET bromodomain protein inhibition reverses chimeric antigen receptor extinction and reinvigorates exhausted T cells in chronic lymphocytic leukemia. *J. Clin. Invest.* **131**, e145459 (2021).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2023

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and, unless otherwise stated, the investigators were not blinded to allocation during experiments and outcome assessment.

Retroviral vector constructs and retroviral production

Plasmids encoding the retroviral vector were prepared using standard molecular biology techniques[45]. LNGFR is a truncated and mutated TNFR family homologue[46], which was used as a control molecule to ensure comparable CAR expression levels from different bicistronic vectors. Synthesis of Rv-1928z, Rv-19BBz and Rv-1928z + 4-1BBL has been previously described[23,47,48]. VSV-G-pseudotyped retroviral supernatants derived from transduced gpg29 fibroblasts (H29) were used to construct stable retroviral-producing cell lines as previously described[49].

AAV targeting for *TRAC*-1928z

The *TRAC* gRNA targets a sequence upstream of the transmembrane domain of the TCRα. This domain is required for TCRα and TCRβ assembly and trafficking to the cell surface. Both non-homologous end joining and integration of the CAR by homology-directed repair (HDR) at this locus would then efficiently disrupt the TCR complex[23]. *TRAC*-1928z is based on the pAAV-GFP backbone (Cell Biolabs). It contains 1.9 kb of genomic *TRAC* flanking the gRNA targeting sequence, a self-cleaving P2A peptide in-frame with the first exon of *TRAC* followed by the 1928z CAR used in clinical trials[24,50].

Isolation and expansion of human T cells

Buffy coats from anonymous healthy donors were purchased from the New York Blood Centre (institutional review board exempted) and peripheral blood was obtained from healthy volunteers. All blood samples were handled following the required ethical and safety procedures. Peripheral blood mononuclear cells were isolated by density gradient centrifugation. T cells were then purified by using the Pan T Cell Isolation kit (Miltenyi Biotec). T cells were stimulated with CD3/CD28 T cell activator Dynabeads (Invitrogen) at 1:1 ratio and cultured in RPMI + 10% FBS, 5 ng ml^{-1} IL-7 and 5 ng ml^{-1} IL-15 (Miltenyi Biotec) for retroviral transduction (Rv-CAR) and gene targeting (*TRAC*-1928z) experiments. The medium was changed every 2 days, and cells were plated at 10^6 cells per millilitre.

Flow cytometry

CAR expression was measured with Alexa-Fluor-647-conjugated goat anti-mouse Fab (115-606-072, Jackson ImmunoResearch). The flow cytometry antibodies used for cell-surface phenotyping are provided in Supplementary Table 1. For intracellular staining, cells were fixed and permeabilized using Foxp3/Transcription Factor staining kit (00-5523-00, eBioscience) according to the manufacturer's protocol. The flow cytometry antibodies used for intracellular studies are provided in Supplementary Table 5. Data were analysed by FlowJo v10.1 (BD). Cell sorting was performed on a BD FACSAria cell sorter. The gating strategies for flow cytometry are provided in Supplementary Fig. 4.

Mouse leukaemia and prostate tumour models

We used 6–12-week-old NOD/SCID/IL-2Rγ null mice (The Jackson Laboratory), under a protocol approved by the Memorial Sloan Kettering Cancer Centre (MSKCC) Institutional Animal Care and Use Committee. All relevant animal use guidelines and ethical regulations were followed. NALM6-expressing firefly luciferase (FFLuc)–GFP cells have been previously described[48]. For the leukaemia model, either male or female mice were inoculated with 5×10^5 FFLuc–GFP NALM6 cells by tail vein injection; CAR T cells were then injected 4 days later at varying doses. For the prostate cancer model, male mice were inoculated with 2×10^6 PC3-PSMA FFLuc–GFP[51] cells by tail vein injection; CAR T cells were then injected 4 weeks later. Both NALM6 and PC3 cells produced similar tumour burdens, and no mice were excluded before treatment. For the *TRAC*-1928z stress test, long-term CAR T cell assessment (Fig. 2) and prostate cancer model (Extended Data Fig. 4), a scrambled gRNA GCACUACCAGAGCUAA CUCA was used as a control. No randomization or blinding methods were used. BLI was performed using the IVIS Imaging System (PerkinElmer) with the Living Image V4.4 software (PerkinElmer) for the acquisition of imaging datasets.

Secondary transplant of $TET2_{bed}$ CAR T cells

A day before the transplant, NSG mice were irradiated with a cumulative dose of 200 cGy. $TET2_{bed}$ CAR T cells (2×10^6) were then injected through the tail vein. For the IL-2 treatment group, mice received 1,000 U of IL-2 twice a week (intraperitoneal). For the IL-7 + IL-15 treatment group, IL-7 was subcutaneously injected at 0.5 µg per mouse per week. IL-15 and IL-15RA were pre-incubated at a 1:6 weight ratio at 37 °C for 30 min before injection (intraperitoneal) in mice at a dose of 2.5 µg (IL-15) + 15 µg (IL-15RA) per week[52]. Mice received exogenous cytokines for 60 days.

Cytotoxicity assays

The cytotoxicity of T cells transduced with a CAR was determined by a luciferase-based assay. NALM6-expressing FFLuc–GFP cells served as target cells. The effector and tumour cells were co-cultured at indicated effector to target (E/T) ratio in the black-walled 96-well plates in triplicate manner with 1×10^5 target cells in a total volume of 100 µl per well. Target cells alone were planted at the same cell density to determine the maximal luciferase expression (relative light units (RLU_{max})). Eighteen hours later, 100 µl luciferase substrate (Bright-Glo, Promega) was directly added to each well. Emitted light was measured by a luminescence plate reader or the Xenogen IVIS Imaging System (Xenogen) with Living Image V4.4 software (Xenogen) for acquisition of imaging datasets. Lysis was determined as $(1 - (RLU_{sample})/(RLU_{max})) \times 100$.

DNA–RNA simultaneous extraction

Cell pellets were resuspended in RLT buffer and nucleic acids were extracted using the AllPrep DNA/RNA Mini Kit (80204, Qiagen) according to the manufacturer's instructions. RNA was eluted in nuclease-free water and DNA in 0.5X buffer EB. Phase separation in cells lysed in TRIzol reagent (15596018, Thermo Fisher) was induced with chloroform. RNA was precipitated with isopropanol and linear acrylamide and washed with 75% ethanol. The samples were resuspended in RNase-free water.

Transcriptome sequencing

After RiboGreen quantification and quality control by Agilent BioAnalyser, 2 ng total RNA with RNA integrity numbers ranging from 7.3 to 9.7 underwent amplification using the SMART-seq v4 Ultra Low Input RNA Kit (63488, Clonetech), with 12 cycles of amplification. Subsequently, 10 ng of amplified cDNA was used to prepare libraries with the KAPA Hyper Prep Kit (KK8504, Kapa Biosystems) using 12 cycles of PCR. Samples were barcoded and run on a HiSeq 4000 in a PE50 run, using the HiSeq 3000/4000 SBS Kit (Illumina). An average of 40 million paired reads were generated per sample and the percent of mRNA bases per sample ranged from 31% to 69%. DESeq2 was used for normalization and differential analysis of transcriptional data.

TCR sequencing

After PicoGreen quantification and quality control by Agilent BioAnalyser, 188–200 ng of genomic DNA was split equally into six reactions and prepared using the immunoSEQ human TCRB Kit (Adaptive Biotechnologies) according to the manufacturer's instructions. In brief, multiplex PCR was used to amplify the CDR3 region for 31 cycles. After clean-up, 2 µl of PCR product was used as input into library preparation with eight cycles of PCR. Barcoded samples were pooled by volume and

sequenced using custom primers on a NextSeq 500 in a SR155 run, using the NextSeq 500/550 Mid Output Kit v2.5 (150 cycles; Illumina). The loading concentration was 1 pM, and 20% spike-in of PhiX was added to the run to increase diversity and for quality control purposes. Raw BCL files were transferred to the immunoSEQ Analyser for processing and analysis.

Exome capture and sequencing

After PicoGreen quantification and quality control by Agilent BioAnalyser, 250 ng of DNA were used to prepare libraries using the KAPA Hyper Prep Kit with eight cycles of PCR. After sample barcoding, 100 ng of library were captured by hybridization using the xGen Exome Research Panel v1.0 (IDT) according to the manufacturer's protocol. PCR amplification of the post-capture libraries was carried out for eight cycles. Samples were run on a HiSeq 4000 in a PE100 run, using the HiSeq 3000/4000 SBS Kit (Illumina). Samples were covered to an average of 111×.

ATAC-seq

Profiling of chromatin was performed by assay for transposase-accessible chromatin using sequencing (ATAC-seq) as previously described[53]. In brief, 50,000 fresh T cells were washed in cold PBS and lysed. The transposition reaction containing TDE1 Tagment DNA Enzyme (20034198, Illumina) was incubated at 37 °C for 30 min. The DNA was cleaned with the MinElute PCR Purification Kit (28004, Qiagen) and material was amplified for five cycles using NEBNext High-Fidelity 2X PCR Master Mix (M0541L, New England Biolabs). After evaluation by real-time PCR, eight additional PCR cycles were done. The final product was cleaned by aMPure XP beads (A63882, Beckman Coulter) at a 1× ratio, and size selection was performed at a 0.5× ratio. Libraries were sequenced on a HiSeq 4000 in a PE50 run, using the HiSeq 3000/4000 SBS Kit (Illumina). An average of 108 million paired reads were generated per sample.

S-EPTS/LM-PCR integration site analysis

Shearing-extension primer tag selection ligation-mediated PCR (S-EPTS/LM-PCR) is a shearing DNA-based integration site analysis method in orientation to the original EPTS/LM-PCR[54]. S-EPTS/LM-PCR starts with shearing of genomic DNA to an intended length of 500 bp using the Covaris M220 instrument. Sheared DNA is split into three equal replicates (500 ng each) and purified, followed by primer extension using two vector, long-terminal-repeat-specific biotinylated primers. The extension product is purified, and biotinylated DNA captured by paramagnetic beads. The captured DNA is ligated to linker cassettes including a molecular barcode, and the ligation product is amplified in an exponential PCR using biotinylated vector-specific and linker-cassette-specific primers. Biotinylated PCR products are magnetically captured, washed and used as template for amplification in a second exponential PCR with barcoded primers, allowing sequencing by MiSeq technology (Illumina). Final preparation for sequencing was done as previously described[55,56]. Applied DNA double barcoding allowed for parallel sequencing of multiple samples in a single sequencing run while minimizing sample cross-contamination. Amplicons were then sequenced on the MiSeq instrument using the V2 Reagent Kit (Illumina).

Integration site computational analysis

Raw sequence data were trimmed according to sequence quality (Phred) and only sequences showing complete identity in both molecular barcodes (linker cassette barcode and sequencing barcodes) were further analysed. An in-house semi-automated bioinformatical data mining pipeline was used to analyse the data[57]. In brief, quality-filtered sequences were trimmed (vector-specific and linker-cassette-specific parts removed) and only sequences that showed at least 18 nucleotides of expected, vector-specific sequence were analysed further to ensure the analysis of true vector–genome junctions. Such trimmed sequences were further filtered in a way that only sequences equal to or larger than 25 bp were aligned to the human genome (UCSC assembly release number hg38, version 3) by Burrows–Wheeler Aligner MEM algorithm (version 0.7.17) for the initial alignment[58]. It was subsequently followed by mapping of potential integration site sequences with BLAST, in which the minimum alignment identity percentage of 95% is used, whereas nearby genes and other integrating features were annotated as previously described according to RefSeq database[59]. The relative sequence count of each detected integration site was calculated in relation to all sequences attributed to the corresponding sample.

Statistical analysis

All statistical analyses were performed using the Prism 9 (GraphPad) software. No statistical methods were used to predetermine sample size. Statistical tests are provided in the figure legends. The Kolmogorov–Smirnov test was used to determine P values in gene set enrichment analysis: $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.0001$ and $^{****}P < 0.00001$.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data generated from RNA-seq and ATAC-seq have been deposited in the Gene Expression Omnibus with the accession number GSE220259. The publicly available datasets used in this study are GSE23321 for central memory and effector memory phenotype comparison, AKL_HTLV1_UP (M7705), AKL_HTLV1_DN (M9815), the angioimmunoblastic T cell lymphoma dataset (GSE6338) and HALLMARK_MYC_V1 (M5926). Source data are provided with this paper.

45. Riviere, I., Brose, K. & Mulligan, R. C. Effects of retroviral vector design on expression of human adenosine deaminase in murine bone marrow transplant recipients engrafted with genetically modified cells. *Proc. Natl Acad. Sci. USA* **92**, 6733–6737 (1995).
46. Gallardo, H. F., Tan, C., Ory, D. & Sadelain, M. Recombinant retroviruses pseudotyped with the vesicular stomatitis virus G glycoprotein mediate both stable gene transfer and pseudotransduction in human peripheral blood lymphocytes. *Blood* **90**, 952–957 (1997).
47. Brentjens, R. J. et al. Eradication of systemic B-cell tumors by genetically targeted human T lymphocytes co-stimulated by CD80 and interleukin-15. *Nat. Med.* **9**, 279–286 (2003).
48. Brentjens, R. J. et al. Genetically targeted T cells eradicate systemic acute lymphoblastic leukemia xenografts. *Clin. Cancer Res.* **13**, 5426–5435 (2007).
49. Gong, M. C. et al. Cancer patient T cells genetically targeted to prostate-specific membrane antigen specifically lyse prostate cancer cells and release cytokines in response to prostate-specific membrane antigen. *Neoplasia* **1**, 123–127 (1999).
50. Brentjens, R. J. et al. CD19-targeted T cells rapidly induce molecular remissions in adults with chemotherapy-refractory acute lymphoblastic leukemia. *Sci. Transl Med.* **5**, 177ra138 (2013).
51. Stephan, M. T. et al. T cell-encoded CD80 and 4-1BBL induce auto- and transcostimulation, resulting in potent tumor rejection. *Nat. Med.* **13**, 1440–1449 (2007).
52. Stoklasek, T. A., Schluns, K. S. & Lefrancois, L. Combined IL-15/IL-15Rα immunotherapy maximizes IL-15 activity in vivo. *J. Immunol.* **177**, 6072–6080 (2006).
53. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
54. Schmidt, M. et al. Detection and direct genomic sequencing of multiple rare unknown flanking DNA in highly complex samples. *Hum. Gene Ther.* **12**, 743–749 (2001).
55. Gabriel, R. et al. Comprehensive genomic access to vector integration in clinical gene therapy. *Nat. Med.* **15**, 1431–1436 (2009).
56. Paruzynski, A. et al. Genome-wide high-throughput integrome analyses by nrLAM-PCR and next-generation sequencing. *Nat. Protoc.* **5**, 1379–1395 (2010).
57. Afzal, S., Wilkening, S., von Kalle, C., Schmidt, M. & Fronza, R. GENE-IS: time-efficient and accurate analysis of viral integration events in large-scale gene therapy data. *Mol. Ther. Nucleic Acids* **6**, 133–139 (2017).
58. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
59. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

Acknowledgements We thank members of the Sadelain laboratory for helpful discussion and feedback; C. Zebley, B. Youngblood, K. Helin and the Sloan Kettering Institute Centre of Epigenetics Research for advice on epigenetic analysis; J. Boyer for western blot support; N.

Socci for advice on exome analysis; M. Schmidt for retroviral integration site analysis; S. Monette and A. Michel from the SKI/CUMC laboratory of Comparative Pathology for conducting pathology analysis; and the following SKI core facilities for their support: Flow Cytometry, Centre of Comparative Medicine and Pathology, Anti-tumour Assessment, Molecular Cytology, Bioinformatics, Integrated Genomics Operation and Cell Therapy and Cell Engineering. Illustrations in Figs. 1a, 2c and 5h and Extended Data Figs. 4a, 7a and 10a were generated using Servier Medical Art. This work was supported by the Pasteur–Weizmann/Servier award, the Leopold Griffuel award, the Leukemia and Lymphoma society (LLS ID: 7014-17) and the MSKCC core grant (P30 CA008748).

Author contributions N.J. and Z.Z. designed the study, performed the experiments, analysed and interpreted data and wrote the manuscript. A.I. and M.L. contributed to RNA-seq and exome analysis. R.K., J.Y. and Y.Z. contributed to ATAC-seq analysis. J.F. and A.D. contributed to animal studies. J.M.-S. contributed to gene targeting. G.G. contributed to vector construction, T cell transduction and animal studies. M.S. designed the study, analysed and interpreted data and wrote the manuscript.

Competing interests The authors declare no competing interests.

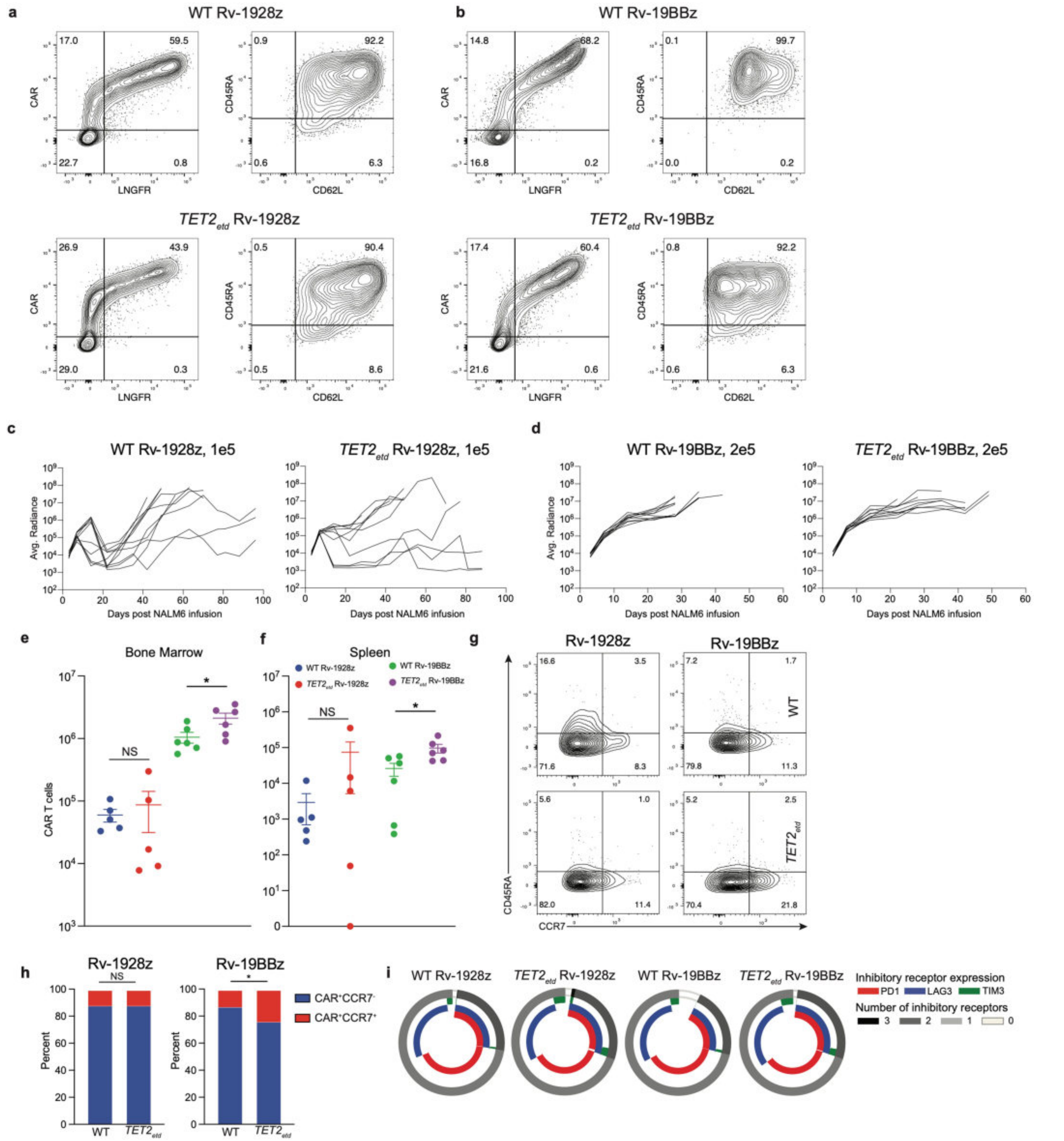
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-05692-z.

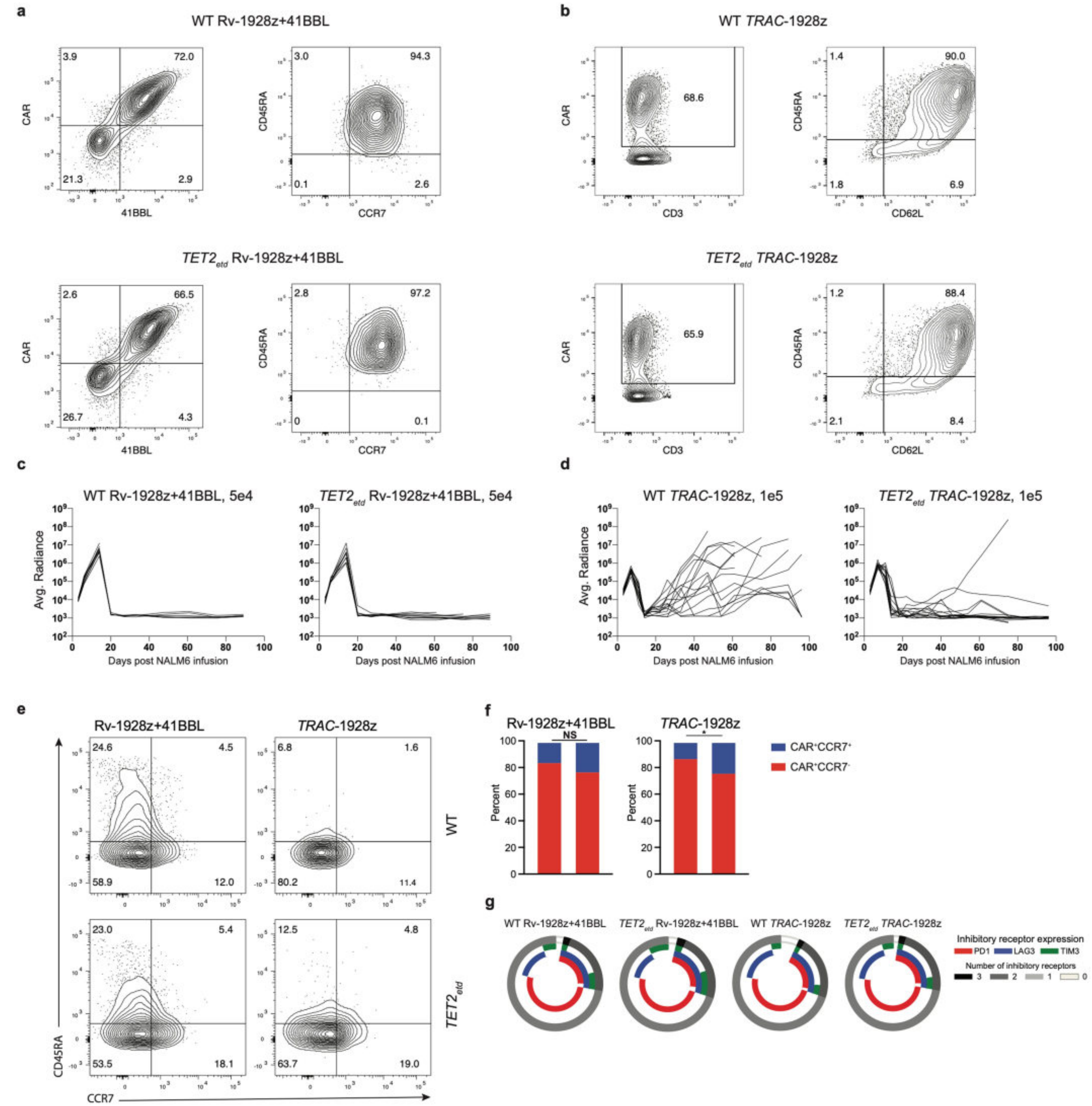
Correspondence and requests for materials should be addressed to Michel Sadelain.

Peer review information *Nature* thanks Stephen Gottschalk and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

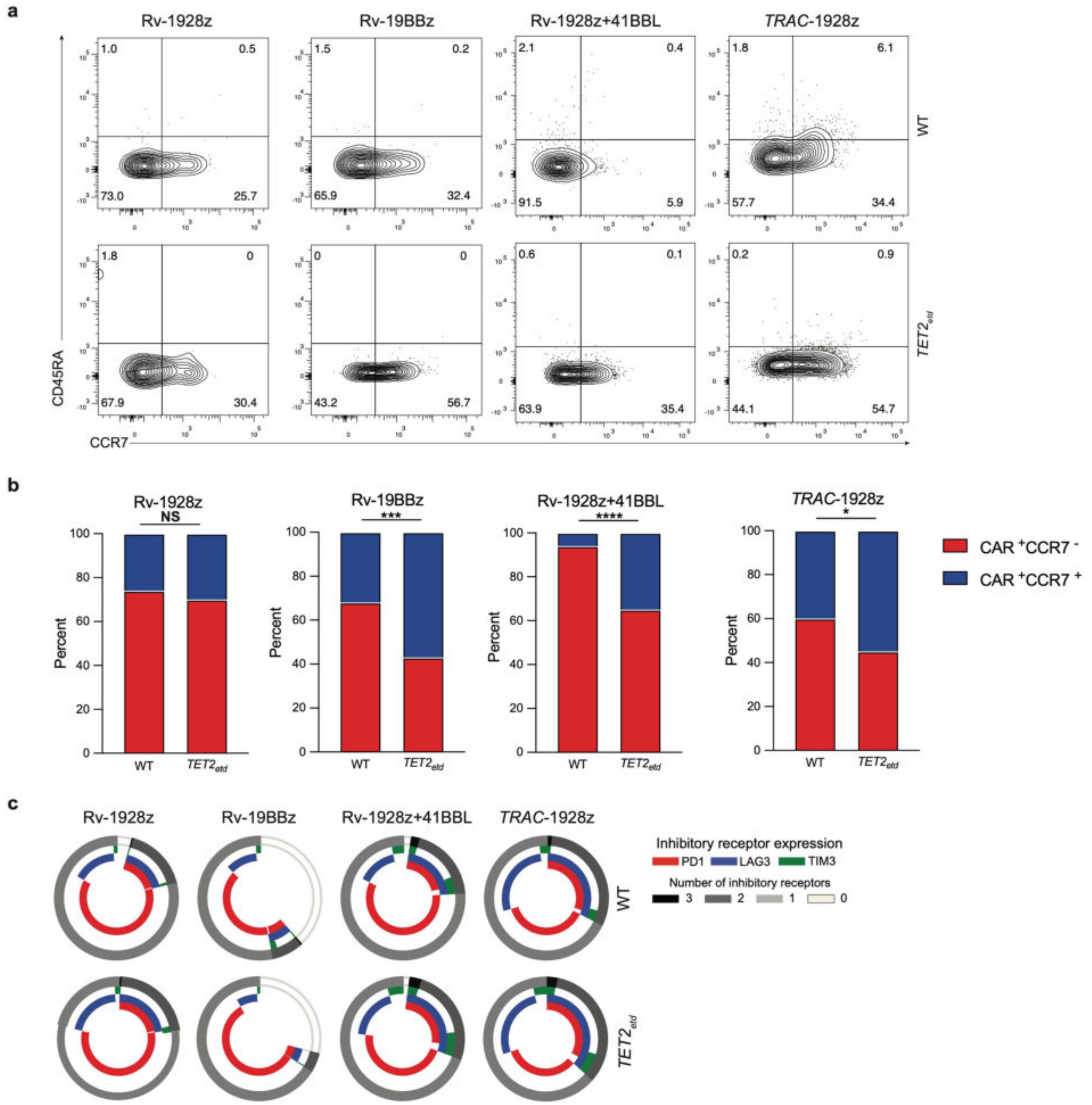
Reprints and permissions information is available at http://www.nature.com/reprints.



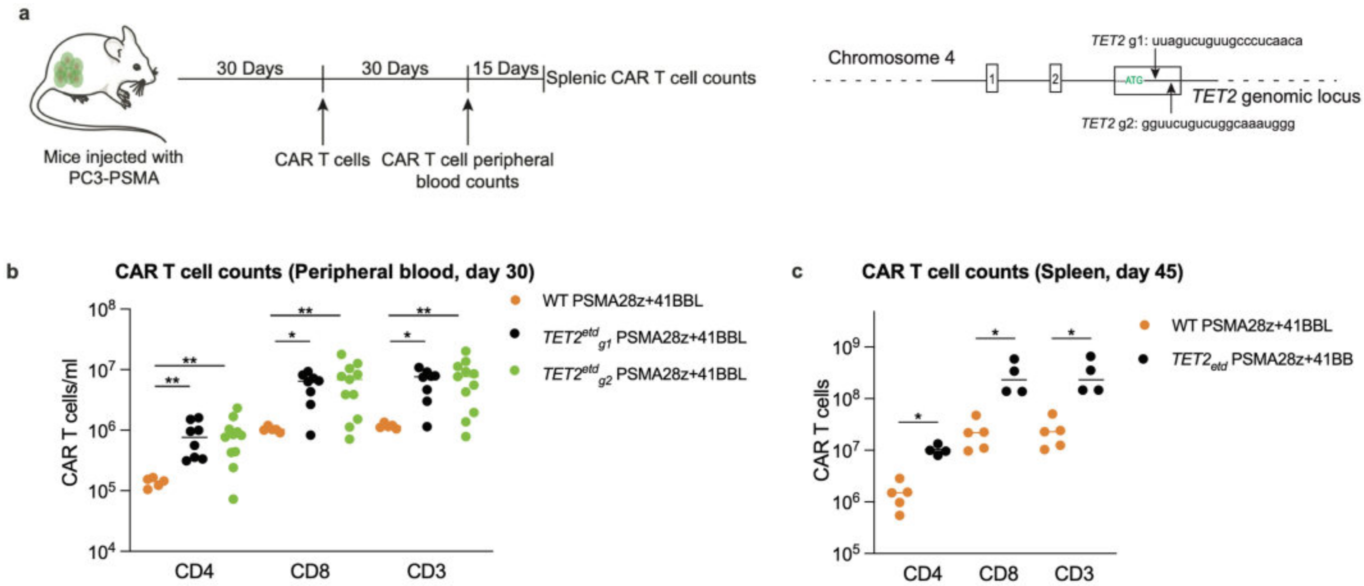
Extended Data Fig. 1 | Rv-1928z and Rv-19BBz pre-infusion and *in vivo* CAR T cell phenotyping. a,**b**, Pre-infusion transduction efficiency and phenotyping by flow cytometry of Rv-1928z (**a**) and Rv-19BBz (**b**) CAR T cells. **c**,**d**, Tumour monitoring of NALM6 bearing mice treated with Rv-1928z (**c**) and Rv-19BBz (**d**) CAR T cells. **e**,**f**, Bone marrow (**e**) and Splenic (**f**) CAR T cell quantification at 3 weeks post infusion. Data is represented as mean±SE [n = 5 (Rv-1928z), n = 6 (Rv-19BBz)]. **g**,**h** Differentiation phenotyping of pooled bone marrow CAR T cells at week 3 post infusion. Data from another experiment included in supplementary information. **i**, CAR T cell inhibitory receptor expression at week 3 post infusion from mouse bone marrow (n = 3). p values were determined by two-sided Mann–Whitney test (**e**,**f**) and two-sided $\chi 2$ test (**h**). $p < 0.05$ was considered statistically significant. p values are denoted: $p > 0.05$, not significant, NS; *, $p < 0.05$. Replicate information for **g**,**i** are available in Supplementary Table 3. Exact p values are available in Supplementary Table 4.



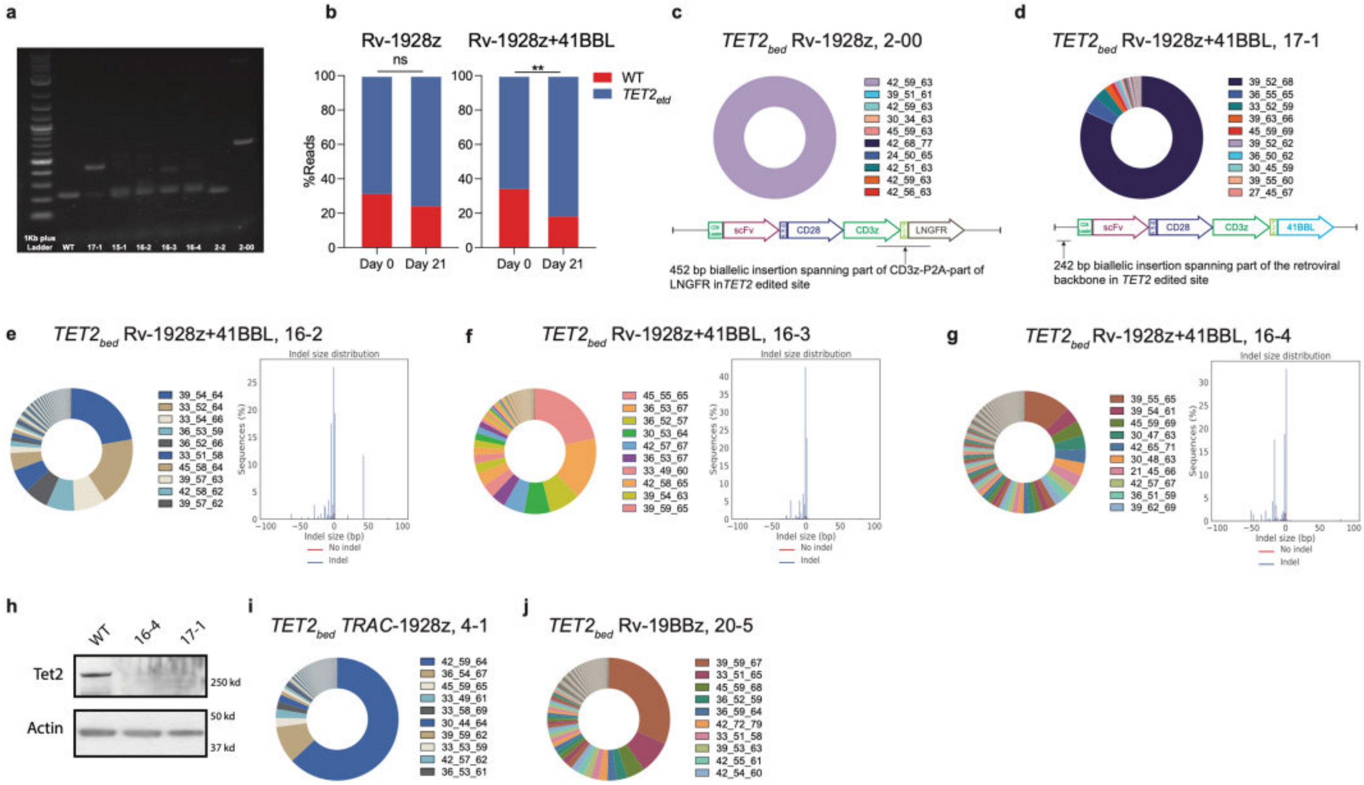
Extended Data Fig. 2 | Rv-1928z+41BBL and *TRAC*-1928z pre-infusion and in vivo CAR T cell phenotyping. **a**,**b**, Pre-infusion transduction efficiency and phenotyping by flow cytometry of Rv-1928z+ 41BBL (**a**) and *TRAC*-1928z (**b**) CAR T cells. **c**,**d**, Tumour monitoring of NALM6 bearing mice treated with Rv-1928z+ 41BBL (**c**) and *TRAC*-1928z (**d**) CAR T cells. **e**,**f**, Differentiation phenotyping of pooled bone marrow CAR T cells at week 3 post infusion. Data from another experiment included in supplementary information. **g**, CAR T cell inhibitory receptor expression at week 3 post infusion from mouse bone marrow (n = 3). p values were determined by two-sided χ2 test (**f**). $p < 0.05$ was considered statistically significant. p values are denoted: $p > 0.05$, not significant, NS; *, $p < 0.05$. Replicate information for **e**,**g** are available in Supplementary Table 3.



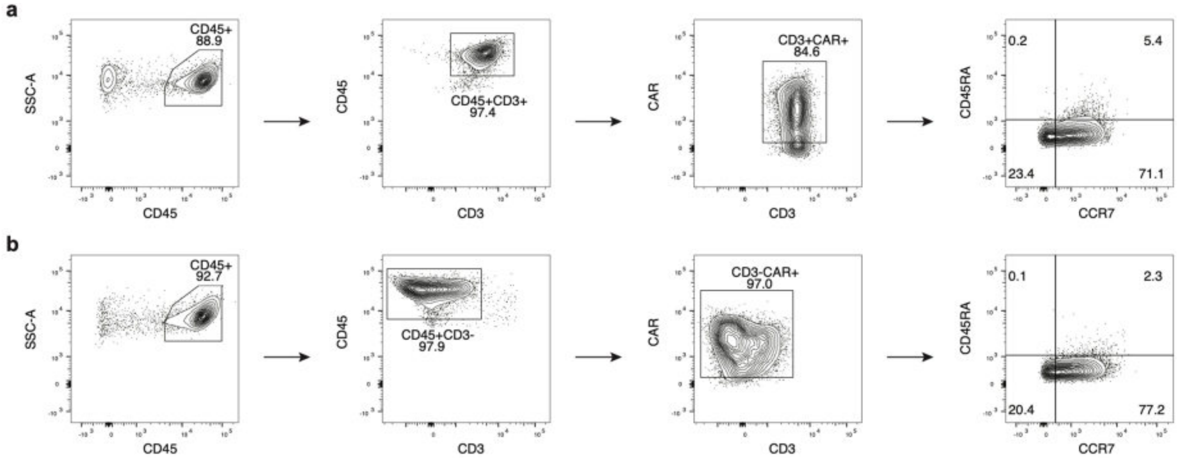
Extended Data Fig. 3 | Long-term CAR T cell phenotypes upon CRISPR/Cas9 editing of *TET2* locus. a,**b**, Differentiation phenotyping of retrovirally encoded CAR T cells (day 90) and *TRAC*-1928z CAR T cells (day 75) isolated from the bone marrow. **c**, Inhibitory receptor expression of bone marrow Rv-1928z, Rv-19BBz, Rv-1928z+41BBL (day 90) and *TRAC*-1928z (day 75) CAR T cells. p values were determined by two-sided $\chi2$ test (**b**). $p < 0.05$ was considered statistically significant. p values are denoted: $p > 0.05$, not significant, NS; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ****, $p < 0.0001$. Replicate information for **a**,**c** are available in Supplementary Table 3.



Extended Data Fig. 4 | Effect of *TET2* editing on CAR T cell accumulation in a prostate cancer model. a, Schematics of the prostate cancer experimental design. *TET2* was edited with the previously discussed gRNA (g1) and an alternative gRNA (g2). PSMA28z+41BBL (PSMA targeted, CD28 costimulated CAR that expresses 41BBL ligand) was used in this study (Dose: 2e5). **b**, CAR T cell counts in the peripheral blood 30 days post infusion of T cells. Bars show median values. **c**, Mice with the top 4 CAR T cell peripheral counts at day 30 across both *TET2* targeting gRNA (g1, n = 2. g2, n = 2) were euthanized at day 45 along with 5 scrambled gRNA treated PSMA28z+41BBL mice and their splenic CAR T cell numbers were quantified. p values were determined by two-sided Mann-Whitney (**b**, **c**) [n = 5 (WT PSMA28z+41BBL), n = 8 ($TET2^{etd}_{g1}$ PSMA29z+41BBL), n = 11 ($TET2^{etd}_{g2}$ PSMA29z+41BBL)]. $p < 0.05$ was considered statistically significant. p values are denoted: *, $p < 0.05$; **, $p < 0.01$. Exact p values are available in Supplementary Table 4. The mouse illustration in part **a** was generated using Servier Medical Art, CC BY 3.0.



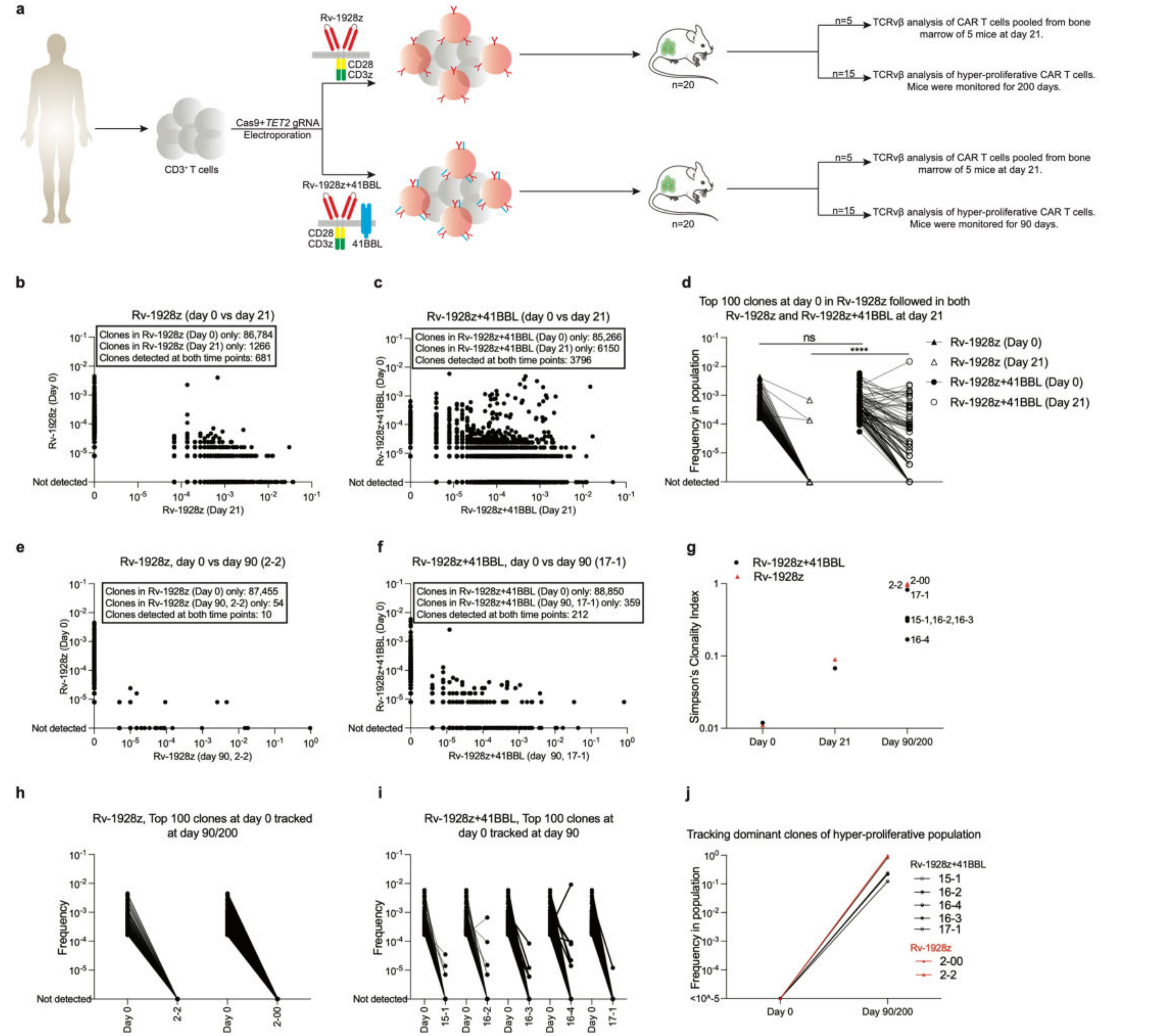
Extended Data Fig. 5 | Clonal expansion in all 4 hyper-proliferative CAR T cell populations. **a**, Gel image of PCR product for WT CAR T cells and hyper-proliferative *TET2*-edited CAR T cells. The PCR is designed to amplify the site of gRNA editing. **b**, Enrichment of *TET2*-editing from pre-infusion (day 0) in mice to day 21 in Rv-1928z and Rv-1928z+41BBL CAR T cells. p values were determined by two-sided χ2 test. **c**,**d**, TCRvβ sequencing reveals hyper-proliferative populations that are dominant for a single clone in $TET2_{bed}$ Rv-1928z (**c**) and Rv-1928z+41BBL (**d**). Part of the retroviral vector that was inserted in the *TET2* alleles of these clones is highlighted in the figures. **e**–**g**, Examples of hyper-proliferative Rv-1928z+41BBL CAR T cell populations that are oligoclonal (left panel) with biallelic *TET2* editing (right panel). **h**, Western blot showing total loss of *TET2* at protein level in different hyper-proliferative populations. **i**,**j**, Examples of oligoclonality in $TET2_{bed}$ *TRAC*-1928z (**i**) and Rv-19BBz (**j**). $p < 0.05$ was considered statistically significant. p values are denoted: $p > 0.05$, not significant, NS; *, $p < 0.05$; **, $p < 0.01$.



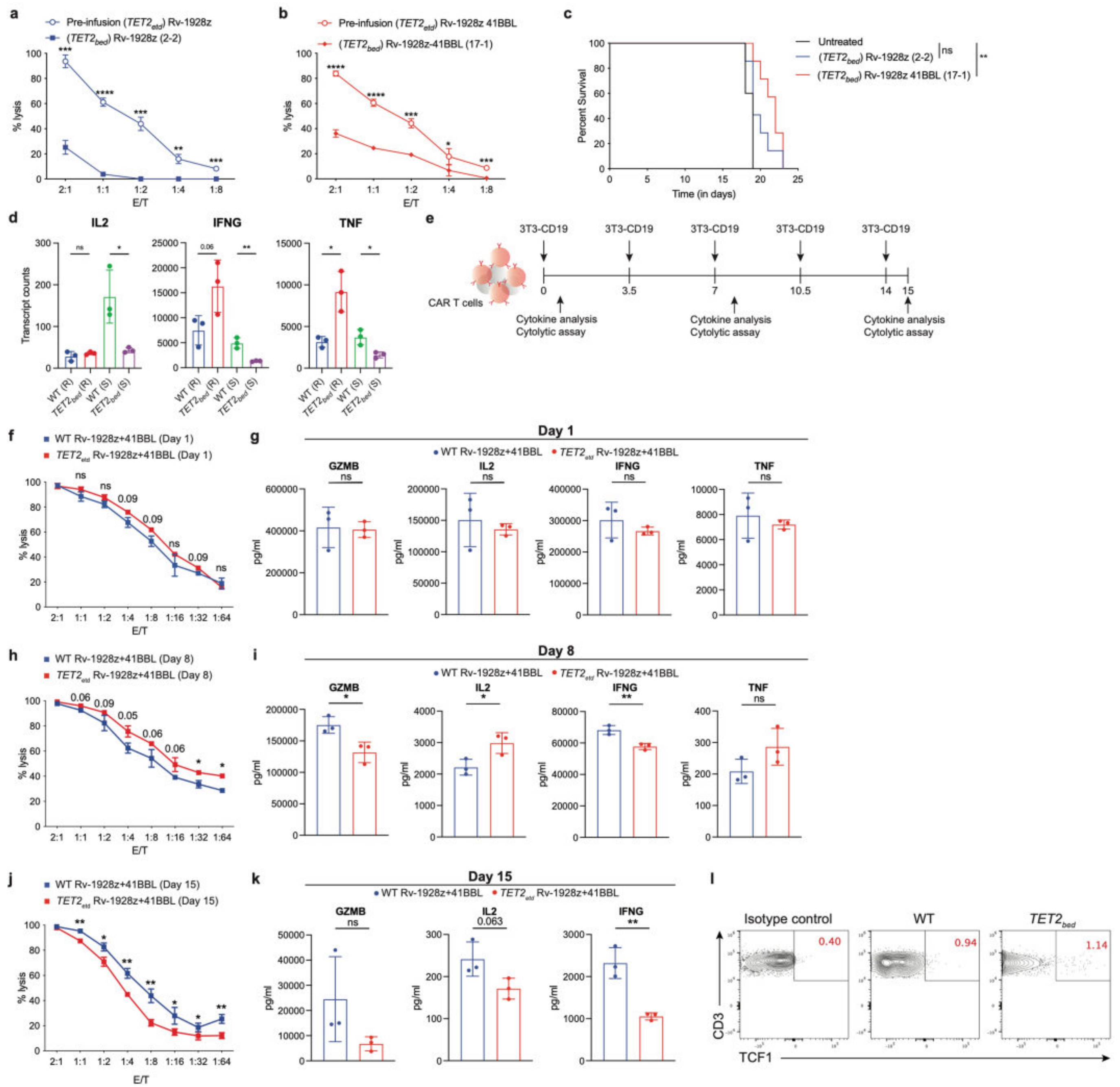
c

CAR T cell	Donor 1	Donor 2
TCR$^+$*TET2*$_{etd}$ Rv-1928z+41BBL	8/10	4/5
TCR$^-$*TET2*$_{etd}$ Rv-1928z+41BBL	7/10	4/5

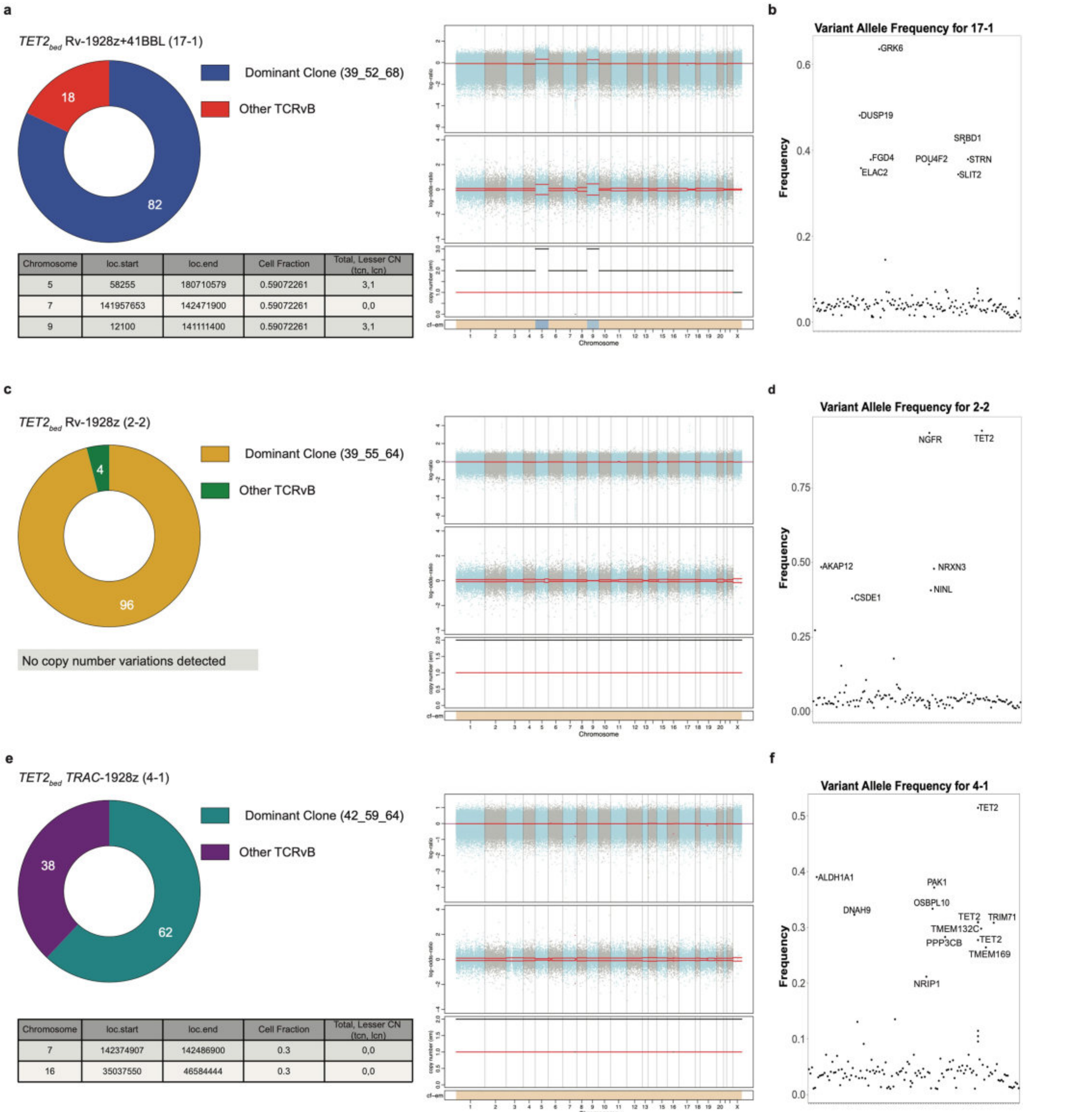
Extended Data Fig. 6 | TCR is dispensable for emergence of hyper-proliferative phenotype in *TET2*-edited Rv-1928z+41BBL CAR T cells. **a**,**b**, Differentiation phenotyping of TCR$^+$*TET2*$_{etd}$ RV-1928z+41BBL (**a**) and TCR$^-$*TET2*$_{etd}$ RV-1928z+41BBL (**b**) CAR T cells. **c**, Summary of emergence of hyper-proliferative phenotype post CAR T cell infusion in mice for different donors. Mice were monitored for 90 days. 2e5 CAR T cells were used for both the groups.



Extended Data Fig. 7 | Properties of the chimeric antigen receptor design determine composition of *TET2*$_{bed}$ hyper-proliferative populations. **a**, Rv-1928z or Rv-1928z+41BBL CAR T cells were generated from the same donor to assess the effect of CAR design on clonal persistence. 5 Mice were euthanized at day 21 to assess clonal diversity post tumour clearance. 15 mice were followed for emergence of a hyper-proliferative phenotype. **b**,**c**, Pair-wise analysis of Rv-1928z (**b**) and Rv-1928z+41BBL (**c**) at day 0 and day 21. **d**, Top 100 Rv-1928z clones at infusion were mapped in the Rv-1928z+41BBL infusion product. These clones were then assessed at day 21 for both the CAR receptors. p values were determined by two-tailed Mann-Whitney test. **e**,**f**, Pair-wise analysis (day 0 vs day 90) of the lone hyper-proliferative population found at day 90 for Rv-1928z CAR receptor (**e**). Representative pair-wise analysis (day 0 vs day 90) of a Rv-1928z+41BBL hyper-proliferative population (**f**). **g**, Changes in clonality index over time in Rv-1928z and Rv-1928z+41BBL CAR T cells. **h**,**i**, Tracking the fate of the 100 most abundant pre-infusion clones in the hyper-proliferative populations of Rv-1928z (**h**) and Rv-1928z+41BBL (**i**). (**j**) Retro-tracking late-stage dominant clones in the infusion product (Day 0). All dominant clones were isolated at day 90 except for 2-00 which was isolated at day 200. $p < 0.05$ was considered statistically significant. p values are denoted: $p > 0.05$, not significant, NS; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ****, $p < 0.0001$. The human, mouse and lipid bilayer illustrations in part **a** were generated using Servier Medical Art, CC BY 3.0.



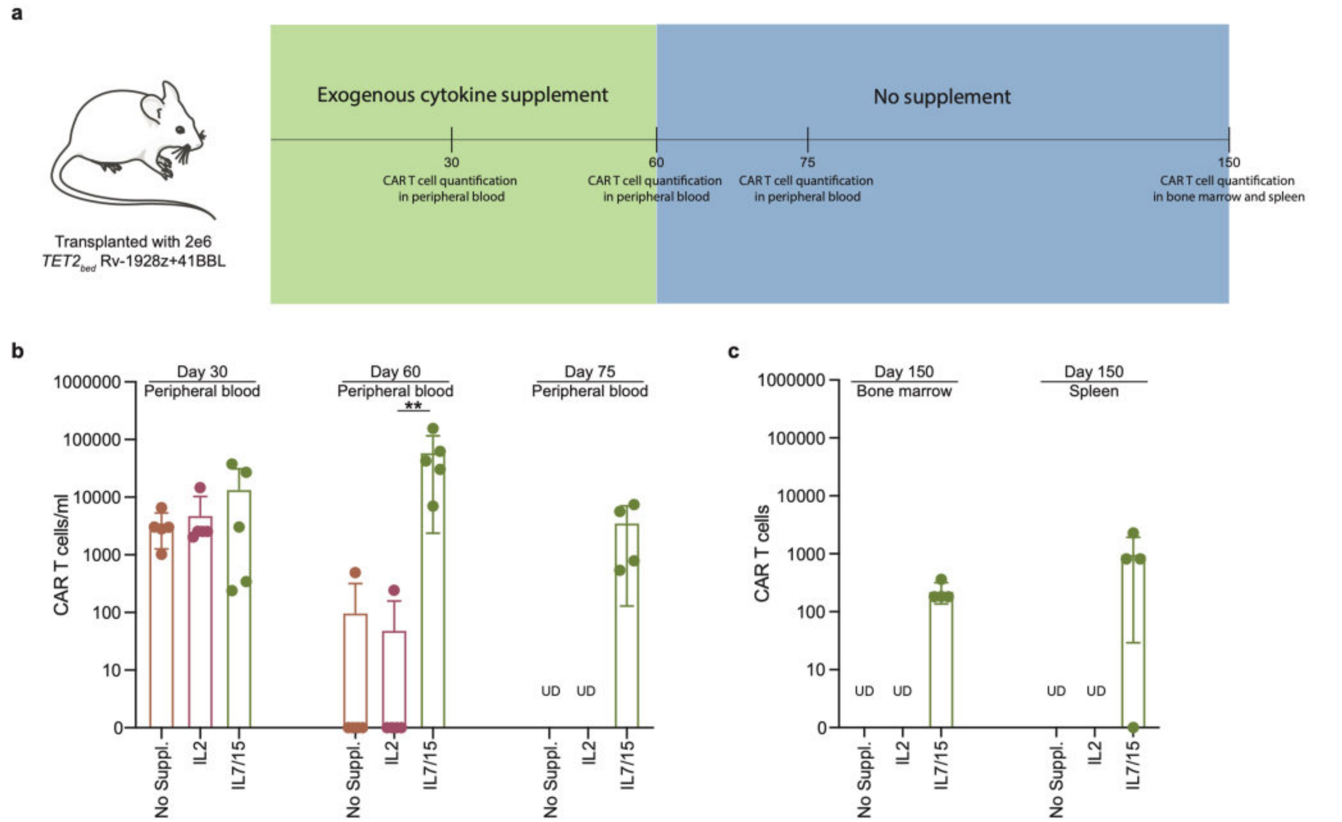
Extended Data Fig. 8 | In vitro and in vivo effector function assessment of *TET2*-edited and hyper-proliferative *TET2$_{bed}$* CAR T cells. a,**b**, In vitro cytolytic activity assessment upon co-culture with NALM6 for 16-h as determined by luciferase activity for pre-infusion *TET2*-edited Rv-1928z (n = 3) and hyper-proliferative *TET2*$_{bed}$ Rv-1928z (2-2) (n = 3) (**a**) and pre-infusion *TET2*-edited Rv-1928z+41BBL (n = 3) and hyper-proliferative *TET2*$_{bed}$ Rv-1928z+41BBL (17-1) (n = 3) (**b**). Data is represented as mean±SD. **c**, NALM6 bearing NSG mice were treated with 2e6 hyper-proliferative *TET2*$_{bed}$ Rv-1928z (n = 7) or *TET2*$_{bed}$ Rv-1928z+41BBL (n = 7) CAR T cells to assess their in vivo anti-tumour efficacy. **d**, Normalized transcript counts of WT Rv-1928z+41BBL and *TET2*$_{bed}$ Rv-1928z+41BBLCAR T cells isolated from mice at day 90. R=Rest (Transcript counts at isolation). S= Stimulated (Transcript counts 24 h post CD3/28 stimulation). Data is represented as mean±SD (n = 3). **e**, Schematic of in vitro repeated rechallenge assay for effector function analysis. **f**,**g**, Day 1 in vitro cytolytic activity assessment (**f**) and effector cytokine assessment (**g**). **h**,**i**, Day 8 in vitro cytolytic activity assessment (**h**) and effector cytokine assessment (**i**). **j**,**k**, Day 15 in vitro cytolytic activity assessment (**j**) and effector cytokine assessment (**k**). Data in **f**–**k** is represented as mean±SD (n = 3). **l**, TCF1 staining of WT Rv-1928z+41BBL and *TET2*$_{bed}$ Rv-1928z+41BBL CAR T cells isolated from mice at day 90. WT samples were a pool of 5 mice. TCF1 staining of other hyper-proliferative *TET2*$_{bed}$ CAR T cells in Supplementary Table 3. p values in **a**,**b**,**f**,**h**,**j** were determined by two-sided Student's unpaired t-test corrected by BKY method. p values in **c** were determined by two-sided Mann-Whitney test. p values in **d**,**g**,**i**,**k** were determined by two-sided unpaired t-test. $p < 0.05$ was considered statistically significant. p values are denoted: $p > 0.1$, not significant, ns. $p < 0.1$ are indicated. *, $p < 0.05$. **, $p < 0.01$. ***, $p < 0.001$. ****, $p < 0.0001$. Exact p values are available in Supplementary Table 4.



Chromosome	loc.start	loc.end	Cell Fraction	Total, Lesser CN (tcn, lcn)
5	58255	180710579	0.59072261	3,1
7	141957653	142471900	0.59072261	0,0
9	12100	141111400	0.59072261	3,1

Chromosome	loc.start	loc.end	Cell Fraction	Total, Lesser CN (tcn, lcn)
7	142374907	142486900	0.3	0,0
16	35037550	46584444	0.3	0,0

Extended Data Fig. 9 | No conserved secondary genetic mutation between different hyper-proliferative *TET2*$_{bed}$ CAR T populations dominant for a single clone. a, (Right panel) Copy number changes in *TET2*$_{bed}$ Rv-1928z+41BBL (17-1). The top panel displays log (ratio) denoted by "(logR)" with chromosomes alternating in the blue and gray. The middle panel displays log (odds-ratio) denoted by "(logOR)". Segment means are plotted in red lines. In the bottom panel total (black) and minor (red) copy number are plotted for each segment. The bottom bar shows the associated cellular fraction (cf). Dark blue indicates high cf. Light blue indicates low cf. Beige indicates a normal segment (total=2, minor=1). The table shows genetic events occurring at >0.1 cf. (Left panel) CAR T cell clonality as determined by vβ sequencing in *TET2*$_{bed}$ Rv-1928z+41BBL (17-1). **b**, Nonsynonymous acquired point mutations in *TET2*$_{bed}$ Rv-1928z+41BBL (17-1). Mutations that occur at a frequency > ((dominant TCRvβ frequency/2) -0.1) or >0.3 whichever is lower is annotated. **c**, (Right panel) Copy number changes in *TET2*$_{bed}$ Rv-1928z (2-2). (Left panel) CAR T cell clonality as determined by vβ sequencing in *TET2*$_{bed}$ Rv-1928z (2-2). **d**, Nonsynonymous acquired point mutations in *TET2*$_{bed}$ Rv-1928z (2-2). **e**, (Right panel) Copy number changes in *TET2*$_{bed}$ *TRAC*-1928z (4-1). (Left panel) CAR T cell clonality as determined by vβ sequencing in *TET2*$_{bed}$ *TRAC*-1928z (4-1). **f**, Nonsynonymous acquired point mutations in *TET2*$_{bed}$ *TRAC*-1928z (4-1).



Extended Data Fig. 10 | Hyper-proliferative $TET2_{bed}$ Rv-1928z+41BBL do not achieve uncontrolled proliferative state upon secondary transplant. **a**, Schematics of secondary transplant of hyper-proliferative $TET2_{bed}$ Rv-1928z+41BBL cells. The exogenous cytokine supplement had to be stopped at day 60 due to deteriorating mice condition in response to frequent injections. **b**, CAR T cell quantification in peripheral blood under different exogenous supplementation at day 30, day 60 and day 75. Each dot represents a mouse. UD: undetected. Data is represented as mean±SD (n = 5). **c**, CAR T cell quantification in bone marrow and spleen at day 150 post CAR T cell infusion. Data is represented as mean±SD (n = 5 for no supplement, and IL2. n = 4 for IL7/15). p values were determined by two-sided Mann–Whitney test (**b**). $p < 0.05$ was considered statistically significant. p values are denoted: $p > 0.05$, not significant, NS; *, $p < 0.05$; **, $p < 0.01$. (**b**). Exact p values are available in Supplementary Table 4. The mouse illustration in part **a** was generated using Servier Medical Art, CC BY 3.0.

nature portfolio

Corresponding author(s): Michel Sadelain

Last updated by author(s): 12/8/2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Xenogen IVIS Imaging System were used to image mice. BD Fortessa or Cytek Aurora cytometers were used to collect flow cytometry data. Tecan Spark microplate reader was used to collect CTL data.
Data analysis	FlowJo (10.1), Living image (V4.4), GraphPad Prism 9, DESeq2 (version1.32.0), Burrows- Wheeler Aligner (version 0.7.17)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data generated from RNAseq and ATACseq experiments has been deposited in GEO (Accession number:GSE220259). The publicly available datasets used in this study are GSE23321 for central memory and effector memory phenotype comparison, AKL_HTLV1_UP (M7705), AKL_HTLV1_DN (M9815), AITL dataset (GSE6338), HALLMARK_MYC_V1 (M5926).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. Sample sizes were estimated based on preliminary experiments.
Data exclusions	Two mice, one from blood quantification (Extended Fig. 4b) and another from splenic quantification (Extended Fig. 4c) were excluded due to technical issue.
Replication	All in vivo tumor efficacy assessments of CAR T cells were performed for at least 2 independent donors. No sample was excluded. The results were consistent between different donors. Across different CAR designs, hyper-proliferative state was observed in all 4 donors that were tested for long-duration (>50 days).
Randomization	Tumor burden was determined by bio-luminescent imaging one day prior to CAR T cell transfer. Since tumour burdens are very even with the NALM6 cell line, no mice were excluded prior to treatment and mice were randomly assigned into treatment groups. For Prostate Cancer mouse model, no mice were excluded prior to treatment and mice were randomly assigned into treatment groups. Buffy coats and human peripheral blood were obtained from anonymous donors to generate human CAR T cells. For comparisons between pre-infusion and hyper-proliferative population, donors were matched to limit confounding variables arising out of intrinsic donor differences. For other experiments, there were no variables to be randomised in this study.
Blinding	Mouse tumour monitoring was performed by an operator who was blinded to treatment groups in addition to the main investigator who was not blind to group allocation. Mouse condition was jointly monitored by the main investigator who was not blind to group allocation and an operator who was blinded to different treatment groups. Exome analysis was performed by an operator who was blinded to different groups. Other experiments could not be blinded due to personnel availability. Blinding is not applicable to data analyses as they are based on objectively measurable data (fluorescence intensity, tumor burden, cell count, transcript and DNA reads).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☐	☒	Antibodies
☐	☒	Eukaryotic cell lines
☒	☐	Palaeontology and archaeology
☐	☒	Animals and other organisms
☐	☒	Human research participants
☒	☐	Clinical data
☒	☐	Dual use research of concern

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☐	☒	Flow cytometry
☒	☐	MRI-based neuroimaging

Antibodies

Antibodies used

For flow cytometry:
Mouse anti-human CD62L BV421, Dilution: 1:100, clone DREG-56, BD, Cat# 563862, Lot: 0016011
Mouse anti-human LAG3 BV605,Dilution: 1:100, clone T47-530, BD, Cat# 745160, Lot: 0057013
Mouse anti-human CD45RA BV605, Dilution 1:100, clone Clone HI100, BD, Cat# 562886, Lot: 0021122
Mouse anti-human CD4 BUV395, Dilution 1:100, Clone SK3, BD, Cat# 563550,Lot: 0188099
Mouse anti-human CD45 BV711, Dilution 1:100, Clone HI30, BD, Cat# 564357, Lot: 9066872
Mouse anti-human CD8 BV510, Dilution 1:100, Clone SK1, BD, Cat# 563919, Lot: 1012142
Mouse anti-human CD3 BUV737, Dilution 1:100, BD, Clone UCHT1, Cat# 612750, Lot: 9212190
Mouse anti-human CD19 BUV496, Dilution 1:100, BD, Clone SJ25C1, CAT# 612938, Lot: 9269743
Mouse anti-human CD271 PE, Dilution 1:100, BD, Clone C40-1457, Cat# 557196,Lot: 8260795
Mouse anti-human CD271 AF647, Dilution 1:100, BD, Clone C40-1457, Cat# 560326, Lot: 7125843

Mouse anti-human PD1 PE, Dilution 1:100, Biolegend, Clone EH12.2H7, Cat# 329906, Lot: B283580
Mouse anti-human Tim3 BV785, Dilution 1:100, Biolegend, Clone F38-2E2, Cat# 345032, Lot: B265346
Mouse anti-human CCR7 PE, Dilution 1:100, Biolegend, Clone G043H7, Cat# 353204, Lot: B281930
Goat Anti-Mouse IgG Antibody AF647, 1:100, Jackson Immuno Research , Polyclonal, Cat# 115-606-072 Lot: 000000154011
Human c-Myc PE-conjugated Antibody, 1:100, R&D System, Clone # 9E10, Cat# IC3696P, Lot ADVP0317111
BATF3 (E3K5H) Rabbit mAb (Alexa Fluor® 647 Conjugate), 1:100, Cell signaling, Clone E3K5H, Cat# 29501S, lot 1
TCF1(TCF7) PE-conjugated antibody, 1:50, Biolegend, Clone 7F11A10, Cat#655208, lot B328728
For Western Blot:
TET2 antibody, 1:100, Cell Signaling, Polyclonal, Cat #45010S, Lot: 1
β-Actin (13E5) Rabbit mAb, 1:2000, Cell Signaling, Clone 13E5, Cat # 4970S, Lot 19

Validation

Antibodies used in this study are commercially available and are validated by the manufacturers, related information are available from their website:
Mouse anti-human CD62L BV421, Cat# 563862, https://www.bdbiosciences.com/en-us/products/reagents/flow-cytometry-reagents/research-reagents/single-color-antibodies-ruo/bv421-mouse-anti-human-cd62l.563862

Mouse anti-human LAG3 BV605,Cat# 745160, https://www.bdbiosciences.com/en-in/products/reagents/flow-cytometry-reagents/research-reagents/single-color-antibodies-ruo/bv605-mouse-anti-human-lag-3-cd223.745160

Mouse anti-human CD45RA BV605, Cat# 562886, https://www.bdbiosciences.com/en-in/products/reagents/flow-cytometry-reagents/research-reagents/single-color-antibodies-ruo/bv605-mouse-anti-human-cd45ra.562886

Mouse anti-human CD4 BUV395, Cat# 563550, https://www.bdbiosciences.com/en-us/products/reagents/flow-cytometry-reagents/research-reagents/single-color-antibodies-ruo/buv395-mouse-anti-human-cd4.563552

Mouse anti-human CD45 BV711, Cat# 564357, https://www.bdbiosciences.com/en-us/products/reagents/flow-cytometry-reagents/research-reagents/single-color-antibodies-ruo/bv711-mouse-anti-human-cd45.564357

Mouse anti-human CD8 BV510, Cat# 563919, https://www.bdbiosciences.com/en-us/products/reagents/flow-cytometry-reagents/research-reagents/single-color-antibodies-ruo/bv510-mouse-anti-human-cd8.563919

Mouse anti-human CD3 BUV737, Cat# 612750, https://www.bdbiosciences.com/en-us/products/reagents/flow-cytometry-reagents/research-reagents/single-color-antibodies-ruo/buv737-mouse-anti-human-cd3.612750

Mouse anti-human CD19 BUV496, CAT# 612938, https://www.bdbiosciences.com/en-us/products/reagents/flow-cytometry-reagents/research-reagents/single-color-antibodies-ruo/buv496-mouse-anti-human-cd19.612938

Mouse anti-human CD271 PE, Cat# 557196, https://www.bdbiosciences.com/en-us/products/reagents/flow-cytometry-reagents/research-reagents/single-color-antibodies-ruo/pe-mouse-anti-human-cd271.557196

Mouse anti-human CD271 AF647, Cat# 560326, https://www.bdbiosciences.com/en-us/products/reagents/flow-cytometry-reagents/research-reagents/single-color-antibodies-ruo/alexa-fluor-647-mouse-anti-human-cd271.560326

Mouse anti-human PD1 PE, Cat# 329906, https://www.biolegend.com/en-us/products/pe-anti-human-cd279-pd-1-antibody-4412?GroupID=BLG5466

Mouse anti-human Tim3 BV785, Cat# 345032, https://www.biolegend.com/en-us/search-results/brilliant-violet-785-anti-human-cd366-tim-3-antibody-11965?GroupID=BLG9937

Mouse anti-human CCR7 PE, Cat# 353204, https://www.biolegend.com/en-us/search-results/pe-anti-human-cd197-ccr7-antibody-7498

Goat Anti-Mouse IgG Antibody AF647, Cat# 115-606-072, https://www.jacksonimmuno.com/catalog/products/115-606-072

Human c-Myc PE-conjugated Antibody, R&D System, https://www.rndsystems.com/products/human-c-myc-pe-conjugated-antibody-9e10_ic3696p

BATF3 (E3K5H) Rabbit mAb (Alexa Fluor® 647 Conjugate), Cell signaling, https://www.cellsignal.com/products/antibody-conjugates/batf3-e3k5h-rabbit-mab-alexa-fluor-647-conjugate/29501?N=4294960176+4294956287&fromPage=plp

TET2 antibody, 1:100, Cell Signaling, Cat #45010S, https://www.cellsignal.com/products/primary-antibodies/tet2-antibody/45010

β-Actin (13E5) Rabbit mAb, Cell Signaling, Cat # 4970S, https://www.cellsignal.com/products/primary-antibodies/b-actin-13e5-rabbit-mab/4970

TCF1(TCF7) PE-conjugated antibody, Biolegend, Cat#655208, https://www.biolegend.com/en-us/lyophilized-control-cells/pe-anti-tcf1-tcf7-antibody-1552

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	NALM6 cell line was obtained from ATCC. PC3 Cell line was obtained from ATCC.
Authentication	COA were provided with cell lines from ATCC. Properties pertinent to the experiments (e.g., GFP or CD19 expression) were confirmed by flow cytometry.
Mycoplasma contamination	All cell lines were tested for mycoplasma contamination and found to be negative.
Commonly misidentified lines (See ICLAC register)	No commonly mis-identified cell lines were used.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Male or female NSG mice, 6-8 weeks old homozygous for the genotype NOD.Cg-Prkdc<scid>Il2rg<tm1Wjl>/SzJ (Jackson laboratory). Animal facility is maintained at a temperatures of 21-23°C, 40-60% humidity and 12-hour light/12-hour dark cycle.
Wild animals	This study did not involved wild animals.
Field-collected samples	This study did not involve samples collected from the field.
Ethics oversight	Memorial Sloan Kettering Cancer Center (MSKCC) Institutional Animal Care and Use Committee.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Buffy coats from anonymous healthy donors were purchased from the New York Blood Centre (institutional review board-exempted) and peripheral blood was obtained from healthy volunteers. All blood samples were handled following the required ethical and safety procedures.
Recruitment	Buffy coats from anonymous healthy donors were purchased from the New York Blood Centre. New York Blood Centre recruits healthy donors under a broad consent covering in vitro laboratory research. Peripheral blood was obtained from healthy volunteers regardless of gender and age.
Ethics oversight	All human blood samples were approved by MSKCC IRB and handled following the required safety procedures. Human buffy coats obatined from New York Blood Centre were IRB-exempt.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

- ☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☒ All plots are contour plots with outliers or pseudocolor plots.
- ☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Sample preparation is described in Material and Methods.
Instrument	Aurora (Cytek Biosciences) Fortessa 3, BD FACSAria Sorter, BD
Software	FlowJo v10.1 (BD)
Cell population abundance	All CAR T cell phenotyping was restricted to CAR positive cells described in supplementary figure 4. RNAseq and ATACseq was performed on sorted CAR positive cells.

Gating strategy

Gating strategies are provided in supplementary figure 4.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Article

Sequence determinant of small RNA production by DICER

https://doi.org/10.1038/s41586-023-05722-4

Received: 30 May 2022

Accepted: 11 January 2023

Published online: 22 February 2023

Check for updates

Young-Yoon Lee[1,2,3], Haedong Kim[1,2,3] & V. Narry Kim[1,2] ✉

RNA silencing relies on specific and efficient processing of double-stranded RNA by Dicer, which yields microRNAs (miRNAs) and small interfering RNAs (siRNAs)[1,2]. However, our current knowledge of the specificity of Dicer is limited to the secondary structures of its substrates: a double-stranded RNA of approximately 22 base pairs with a 2-nucleotide 3′ overhang and a terminal loop[3–11]. Here we found evidence pointing to an additional sequence-dependent determinant beyond these structural properties. To systematically interrogate the features of precursor miRNAs (pre-miRNAs), we carried out massively parallel assays with pre-miRNA variants and human DICER (also known as DICER1). Our analyses revealed a deeply conserved *cis*-acting element, termed the 'GYM motif' (paired G, paired pyrimidine and mismatched C or A), near the cleavage site. The GYM motif promotes processing at a specific position and can override the previously identified 'ruler'-like counting mechanisms from the 5′ and 3′ ends of pre-miRNA[3–6]. Consistently, integrating this motif into short hairpin RNA or Dicer-substrate siRNA potentiates RNA interference. Furthermore, we find that the C-terminal double-stranded RNA-binding domain (dsRBD) of DICER recognizes the GYM motif. Alterations in the dsRBD reduce processing and change cleavage sites in a motif-dependent fashion, affecting the miRNA repertoire in cells. In particular, the cancer-associated R1855L substitution in the dsRBD strongly impairs GYM motif recognition. This study uncovers an ancient principle of substrate recognition by metazoan Dicer and implicates its potential in the design of RNA therapeutics.

Dicer, a multidomain ribonuclease (RNase) III, serves as a key player in RNA silencing by cleaving double-stranded RNA (dsRNA) into small RNAs of 21–25 nucleotides (nt) in length[1,2]. Endogenous siRNAs and miRNAs are produced from long RNA duplexes and RNA hairpins, respectively. The miRNA pathway in metazoan species involves another RNase III, Drosha, that cleaves primary miRNA (pri-miRNA) transcript to release pre-miRNA, a hairpin of about 70 nt with a characteristic 2-nt 3′ overhang[12–16]. Dicer cleaves pre-miRNA to produce a duplex of about 22 nt with a 2-nt 3′ overhang at both ends[2]. After loading onto the Argonaute (Ago) protein, one strand of the duplex remains as a mature miRNA that functions as a guide to base-pair with the cognate target[17,18]. Targeting specificity relies on the precision of processing because even a small change in the processing site can alter the 'seed' sequence (2–7-nt region relative to the 5′ end of the guide RNA) critical for target binding[19–21].

In human, DICER is known to recognize its substrates by relying solely on their secondary structural features, such as the 2-nt 3′ overhang, a dsRNA stem of about 22 base pairs (bp) and a terminal loop[3–11]. According to the current model, DICER acts as a 'molecular ruler' that measures 22 nt from the ends of pre-miRNA[3–6] (Extended Data Fig. 1a). The 3′ end is recognized by a conserved '3′ pocket' in the PAZ domain of DICER. Some DICER homologues also have a '5′ pocket' in the platform domain to capture the 5′ end[5,22,23]. These pockets anchor the ends most effectively when the ends are in a 2-nt 3′-overhang arrangement[3–5,24]. As the catalytic centre of DICER is separated by a fixed distance from these pockets, DICER can measure a specified length (22 nt in human) from the 5′ end ('5′ counting rule') and the 3′ end ('3′ counting rule')[3–6]. The relative contribution of the 5′ and 3′ counting is influenced by thermodynamic stability because an unstable 5′ end can be readily frayed and inserted into the 5′ pocket, facilitating the 5′ counting mechanism[5].

As the termini of pre-miRNAs are created by DROSHA, DICER is considered to play a passive role when it comes to the determination of miRNA sequences (Extended Data Fig. 1a). For instance, pre-let-7a-1 is cleaved essentially at a single site counted from the 5′ end (Extended Data Fig. 1b)[5]. However, some pre-miRNAs do undergo alternative processing at the DICER level[25,26]. Most notably, pre-miR-324 is uridylated frequently at the 3′ end, which results in a shift of the DICER cleavage site[27] (Extended Data Fig. 1b).

We noticed that the end counting rules cannot fully explain the processing pattern of pre-miR-324 or its variants (Extended Data Fig. 1b,c), suggesting that there may be a yet-unknown mechanism by which DICER engages in processing. For example, when we used a pre-miR-324-derived substrate with a symmetric stem and a 3-nt 3′ overhang, it was cleaved at three sites (Extended Data Fig. 1c, lane 3).

[1]Center for RNA Research, Institute for Basic Science (IBS), Seoul, Republic of Korea. [2]School of Biological Sciences, Seoul National University, Seoul, Republic of Korea. [3]These authors contributed equally: Young-Yoon Lee, Haedong Kim. ✉e-mail: narrykim@snu.ac.kr

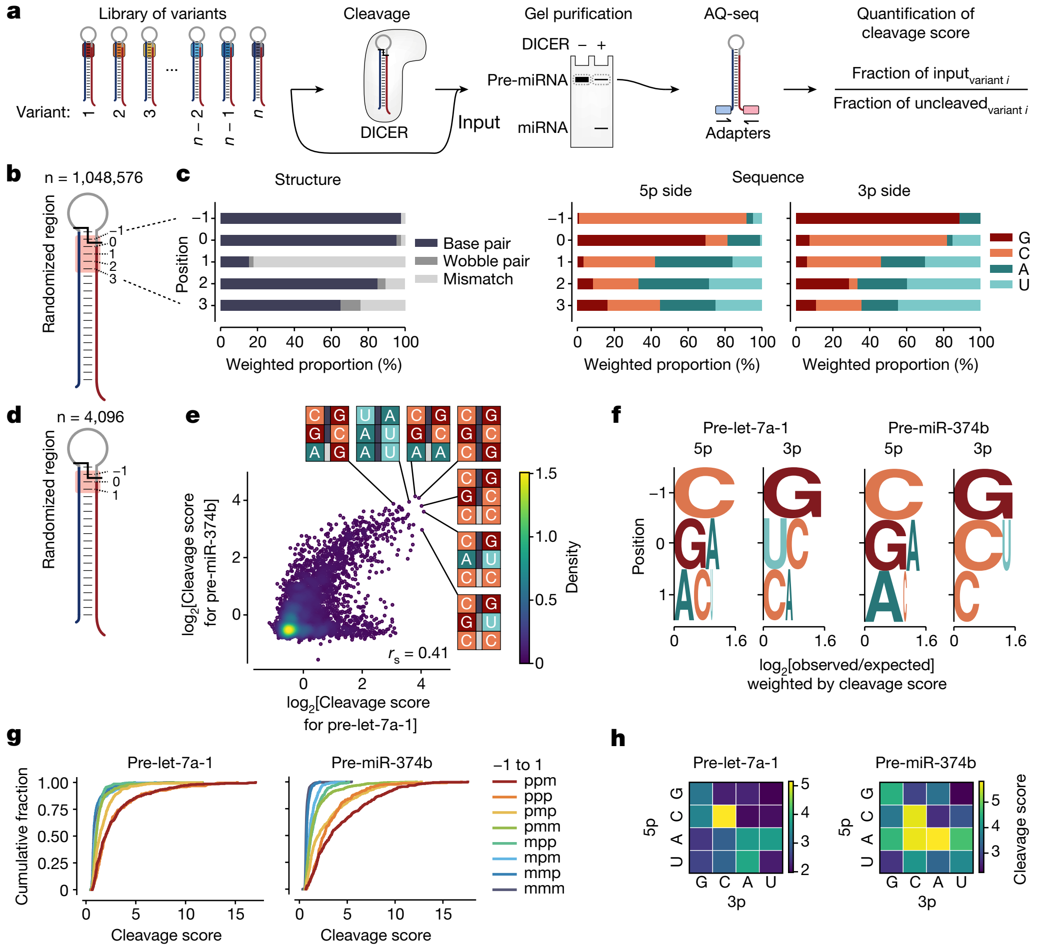


Fig. 1 | Massively parallel assay identifies the GYM motif. a, Schematic outline of the massively parallel assay. **b**, Substrate design for the first screening ($n = 1{,}048{,}576$ variants). Sequence variation was introduced to five base pairs at the positions −1 to 3 relative to the starting position of 3p miRNA. **c**, Massively parallel assay results from the first screening. Structural (left) and sequence (middle and right) preferences of DICER substrates at the indicated positions are shown. Proportions of the features within the top 0.1% of variants were weighted by their cleavage scores. Variants with more than 100 read counts in the input were included in this analysis. **d**, Substrate design for the second screening ($n = 4{,}096$ variants). Sequence variation was introduced to three base pairs at the positions −1 to 1. **e**, Distribution of the cleavage scores measured from the second screening. Top five motifs in each backbone are shown. Analyses were carried out using data from the condition for which about 20% of substrates were cleaved, which is true also for **f**–**h**. r_s, Spearman correlation coefficient. **f**, Enrichment of sequences of the top 1% variants. Expected proportions were calculated by assuming that the four bases have equal contributions to the total cleavage score of the top 1% variants. **g**, Structural impact on cleavage scores. G–U pair was considered as a mismatch only when it is in between mismatches. p, pair; m, mismatch. **h**, Impact of the base combinations at position 1 on cleavage scores. Variants with base pairs at all but position 1 were included in this analysis.

The 5′ and 3′ counting rules explained products A and B, respectively, but not product C. Another variant with an alteration near the cleavage site failed to generate product C (Extended Data Fig. 1c, lane 4), suggesting that there may be a critical element near the cleavage site. As previous studies focused only on secondary structures outside this region and have not investigated the substrate specificity in a comprehensive manner, it remains unknown whether and how DICER recognizes its substrates in a sequence-specific manner.

Identification of a sequence motif

To comprehensively interrogate the upper stem region, we implemented a massively parallel assay that enables quantitative testing of a large number of variants (Fig. 1a). In brief, we synthesized 1,048,576 pre-let-7a-1 variants by randomizing the sequences within a 5-bp window in the upper stem (−1 to +3 relative to the cleavage site in the 3p strand) that is predicted to contact DICER in our structural model (Fig. 1b and Extended Data Fig. 2a,b). After a brief incubation with purified human DICER protein (Extended Data Fig. 2c,d), in which 5% of the substrate pool was cleaved, the uncleaved RNAs were gel-purified and sequenced by the accurate quantification by sequencing (AQ-seq) method that allows efficient ligation of structured RNAs[25,28]. The processing efficiency was quantified by dividing the fraction of each variant in the input population by that of the uncleaved reads after reaction (Fig. 1a). We refer to this metric as the ‘cleavage score’.

To identify the determinants for processing, we examined the pre-miRNA variants scoring within the top 0.1%. These top variants showed an overall tendency for base-pairing as expected (Fig. 1c, left panel). However, we found that a mismatch was enriched at position 1. In addition to this structural feature, we found some sequence preferences (Fig. 1c, middle and right panels). At positions −1 and 0, the 5′-C–G-3′ pair and 5′-G–C-3′ pair were strongly favoured, respectively. At position 1, G is depleted and C is enriched on both strands. Sequence preference was less prominent at positions 2 and 3.

To increase the sequencing depth, we carried out a second round of massively parallel assays, with two 3-bp windows (randomizing positions −1 to 1 and 1 to 3) to generate 4,096 variants per window (Fig. 1d and Extended Data Fig. 2b). In addition, we used another pre-miRNA

backbone, pre-miR-374b, that has a relatively simple structure (Extended Data Fig. 2b). The cleavage reaction was carried out briefly under conditions for which 10%, 20% or 30% of the substrate pool was cleaved (Extended Data Fig. 3a–d). Cleavage scores increased with incubation time but showed strong correlations between conditions (Extended Data Fig. 3e–h), indicating that the experiments were within dynamic ranges.

The variants in the window −1 to 1 showed a broad distribution in their cleavage scores (Fig. 1e). Two different pre-miRNAs showed clear correlations, particularly among the top scoring sequences. These results indicate that the region −1 to 1 may play an important role in the control of pre-miRNA processing, independently of the backbone. With the other window randomizing the region 1 to 3, the cleavage scores did not show large differences between the variants, suggesting that this region is less impactful than the region −1 to 1 (Extended Data Fig. 3i).

In line with the initial screening data, we could identify sequence preferences at positions −1, 0 and 1 (Fig. 1f). At position −1, the 5′-C–G-3′ pair is strongly enriched. At position 0, purines (mainly G) are enriched in the 5p strand and pyrimidines prevail in the 3p strand, forming a base pair. At position 1, A or C is favoured and G or U is depleted.

We also found that base-pairing is overall beneficial to processing, but a mismatch at position 1 is associated with high cleavage scores (Fig. 1g and Extended Data Fig. 4a,b). Some base combinations at position 1 are favoured over others, with the C–C mismatch generally exerting the strongest effects (Fig. 1h and Extended Data Fig. 4c,d). These sequence and structure preferences at position 1 were detectable in almost all contexts examined, regardless of the sliding windows, backbones and reaction times.

Collectively, the findings of this analysis revealed a position-dependent 3-bp motif that strongly promotes processing: a paired G, a paired pyrimidine (Y), and a mismatched C or less favourably A (M) at positions −1, 0 and 1 on the 3p side, respectively (with DICER cleaving between G and Y). We therefore named this element the GYM motif. It is important to note that although there is a clear consensus sequence for the GYM motif, the cleavage scores of the variants are actually a continuum. Thus, the GYM motif should be perceived as a quantitative feature (based on the cleavage scores) rather than being defined in a binary manner (based on the sequence consensus). Therefore, we devised a metric termed the ‘GYM score’ to indicate the strength of the motif. The GYM score was assigned to each 3-bp variant by averaging its cleavage scores measured in the contexts of pre-let-7a-1 and pre-miR-374b with 20% cleavage, and by normalizing the average (0–100).

The GYM motif enhances dsRNA processing

To validate the high-throughput data, we carried out in vitro processing assays using pre-let-7a-1 with various GYM motifs (Fig. 2a). A representative variant with a high-score motif (GCm, GYM score 94) was processed more efficiently than the wild type (GYM score 43; Fig. 2a and Extended Data Fig. 5a,b). We further tested a variant with a G–C pair at position 1 (GCp, scoring 38) and that with a UA dinucleotide at positions −1 and 0 (UAm, scoring 30) that were processed less efficiently than the GCm variant. Finally, the variant replacing both parts (UAp, scoring 14) performed least efficiently, indicating that both the paired GY and the mismatched M parts are important and confer additive effects. We made similar observations in the presence of TRBP, a dsRNA-binding protein that associates with DICER, suggesting that TRBP does not influence the recognition of the GYM by DICER (Extended Data Fig. 5c).

Next, to examine the activity of the motif in cells, the GYM variants were incorporated into an artificial pri-miRNA hairpin[29] embedded in the 3′ untranslated region of a luciferase reporter construct (Extended Data Fig. 5d). This construct has been used to monitor the activity of DROSHA because the hairpin cleavage reduces luciferase expression[30]. The luciferase level was not changed by the GYM motif, indicating that the motif does not impact DROSHA processing. By contrast, the mature miRNA levels from these constructs correlated with the GYM score (Extended Data Fig. 5d). This is consistent with the above in vitro results, and further shows that the GYM motif operates regardless of the backbone.

We next questioned whether this motif also contributes to cleavage site selection. We designed pre-miRNA-like duplexes with the high-score GYM motif either properly positioned or relocated by 1 bp towards the terminus. The relocation resulted in a cleavage site shift by 1 bp (Fig. 2b). The motif lacking the GC dinucleotide (UAm) also affected the cleavage site selection, albeit moderately. By contrast, the motif lacking the mismatch (GCp) did not notably alter the cleavage site. Therefore, the main impact of the GYM motif on cleavage site selection is exerted by the mismatched M, and the GC dinucleotide reinforces the influence of the mismatch and increases processing efficiency (Extended Data Fig. 5e). Collectively, these results indicate that the GYM motif not only promotes efficient processing, but also serves as an important determinant of cleavage site.

To gain further insights into the functional relevance of the GYM motif in other species, we examined *Drosophila* Dicer-1 (Dcr-1) using our substrates. Like human DICER, fly Dcr-1 exhibited a preference for the high-score GYM motif (Fig. 2c and Extended Data Fig. 5f). Moreover, repositioning the motif altered the cleavage site of Dcr-1 (Fig. 2d). Thus, the GYM motif may play a deeply conserved role in pre-miRNA processing in metazoan species.

The current model of dsRNA processing posits that DICER follows the 5′ counting and 3′ counting rules[3–6]. To investigate how the GYM motif interplays with these end counting rules, we prepared a series of RNA duplexes with a low-score GYM motif (UAp), which shows a mixture of products from 5′ counting and 3′ counting according to the varying 1–3-nt 3′ overhang (Fig. 2e, purple and green arrowheads, respectively). In the presence of a high-score GYM motif (GCm), a single site was chosen, yielding a homogeneous product, regardless of the overhang length (Fig. 2e, blue arrowheads). Similarly, in another set of substrates with a terminal base pair (to promote 3′ counting), the high-score GYM motif not only increased the cleavage efficiency but also abolished the 3′ end counting (Extended Data Fig. 5g). Thus, the optimal GYM motif is a dominating determinant of cleavage site decision, overriding the 5′ and 3′ counting rules when in conflict.

The dsRBD of DICER recognizes the motif

To gain mechanistic insights into the recognition of the GYM motif, we built a structural model of the DICER–dsRNA complex in a dicing state. In this predicted structure, the C-terminal dsRBD is located near the GYM motif, suggesting a role for dsRBD (Fig. 2f). Indeed, a DICER mutant lacking the dsRBD (ΔdsRBD) cannot distinguish a mismatch at position 1 (Fig. 2g and Extended Data Fig. 6a). Contrariwise, the GC dinucleotide at positions −1 and 0 was still favoured over the UA dinucleotide. Thus, the dsRBD has a critical role in sensing the mismatch, whereas bases at positions −1 and/or 0 may be recognized by other domain(s).

Next, we examined cleavage site selection and found that, unlike wild-type DICER, ΔdsRBD is no longer influenced by the position of the GYM motif (compare Fig. 2b with Fig. 2h). These results support the role of the dsRBD in the recognition of the GYM motif for cleavage site decision.

These results prompted a closer inspection of the dsRBD–dsRNA interface to identify the residues responsible for the motif recognition. Among the amino acids lining the first α-helix of the dsRBD in the minor groove of dsRNA, a highly conserved arginine residue (R1855) is located in very close proximity to the base at position 1 (Fig. 2f and Extended Data Fig. 6b). A point mutation resulting in the substitution of this arginine residue to leucine (R1855L) was found in cancer through The Cancer Genome Atlas[31,32], implying its defective role in pre-miRNA processing. Indeed, alterations of this arginine to

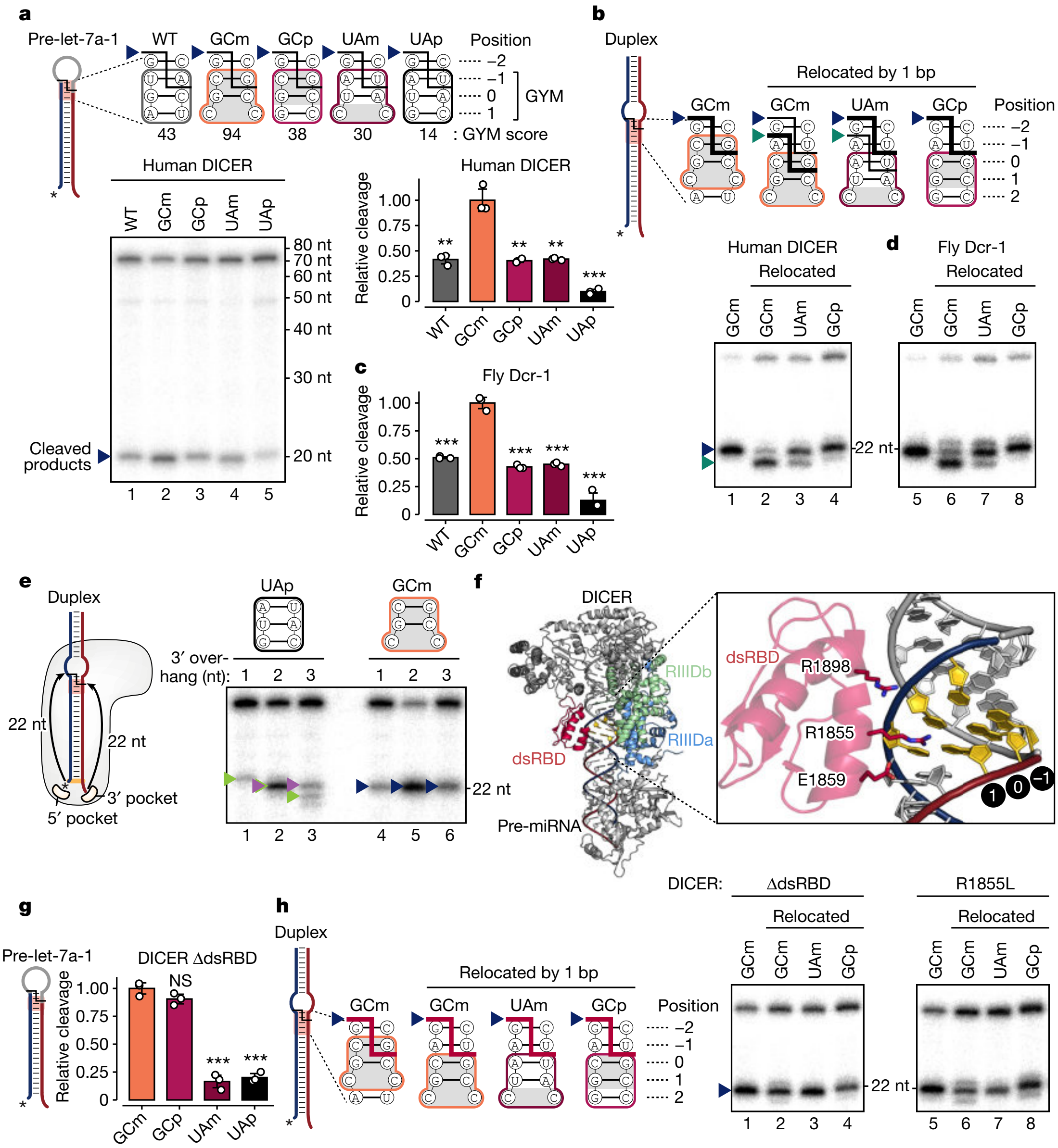


Fig. 2 | The GYM motif recognized by the dsRBD enhances pre-miRNA processing and determines the cleavage site. a, In vitro processing of pre-let-7a-1 variants by human DICER. RNA substrates vary in their sequences at positions −1 to 1 (top). Pre-let-7a-1 variants were processed by human DICER and resolved on a urea–polyacrylamide gel (left). Relative cleavage was calculated by quantifying the band intensity (1 − uncleaved/input) (right). Bars indicate mean ± s.d. ($n = 3$, independent experiments). $^{**}P < 0.01$, $^{***}P < 0.001$ by two-sided Student's *t*-test compared to GCm. WT, wild type. **b**, In vitro processing of duplex variants by human DICER. RNA substrates vary in their sequences at positions −1 to 2 (top). Duplex variants were processed by human DICER and resolved on a urea–polyacrylamide gel (bottom). Cleavage products and their corresponding cleavage sites are marked with arrowheads. **c**, In vitro processing of pre-let-7a-1 variants by fly Dcr-1. Bars indicate mean ± s.d. ($n = 3$, independent experiments). $^{***}P < 0.001$ by two-sided Student's *t*-test compared to GCm. **d**, In vitro processing of duplex variants by fly Dcr-1. **e**, In vitro processing of duplex variants by human DICER. The duplexes had a mismatch at the terminus (marked in orange) so that the 5′ end can be incorporated into the 5′ pocket (left). Duplex variants were processed by human DICER and resolved on a urea–polyacrylamide gel (right). The products from 5′ counting, 3′ counting and the GYM motif are indicated with purple, green, and dark blue arrowheads, respectively. **f**, A structural model of human DICER in a dicing state. The structures of human DICER with a pre-miRNA substrate in a pre-dicing state (PDB: 5ZAL)[23] and *Aquifex aeolicus* RNase III with a dsRNA substrate (PDB: 2EZ6)[39] were used for structural modelling (left). RNase III a (RIIIDa) and RNase III b (RIIIDb) are shown in blue and green, respectively. The dsRBD (marked in red) is in the vicinity of the GYM motif (marked in yellow) (right). **g**, In vitro processing of pre-let-7a-1 variants by DICER(ΔdsRBD). Bars indicate mean ± s.d. ($n = 3$, independent experiments). $^{***}P < 0.001$ by two-sided Student's *t*-test compared to GCm; NS, not significant. **h**, In vitro processing of duplex variants by either the DICER(ΔdsRBD) or DICER(AA) mutant. RNA substrates vary in their sequences at positions −1 to 2 (left). Duplex variants were processed by human DICER and resolved on a urea–polyacrylamide gel (right). The cleavage product and its corresponding cleavage site marked with the arrowhead are largely unaffected by the GYM motif variations, which contrasts the result from WT DICER shown in **b**,**d**. *, radiolabelled 5′ phosphate. For gel source data, see Supplementary Fig. 1.

leucine (R1855L) or alanine (R1855A) were both sufficient to reduce the site-specific recognition of the mismatch in the GYM motif (Extended Data Fig. 6c). We further interrogated the other two highly conserved residues—E1859 and R1898—situated near the mismatch and found that E1859 also contributes to the motif recognition (Fig. 2f and Extended Data Fig. 6b,c). The R1855L mutant and the double mutant (R1855A/E1859A, or 'AA') lost their ability to favour a mismatch at position 1 (for efficiency, compare Fig. 2a with Extended Data Fig. 6d,e; for

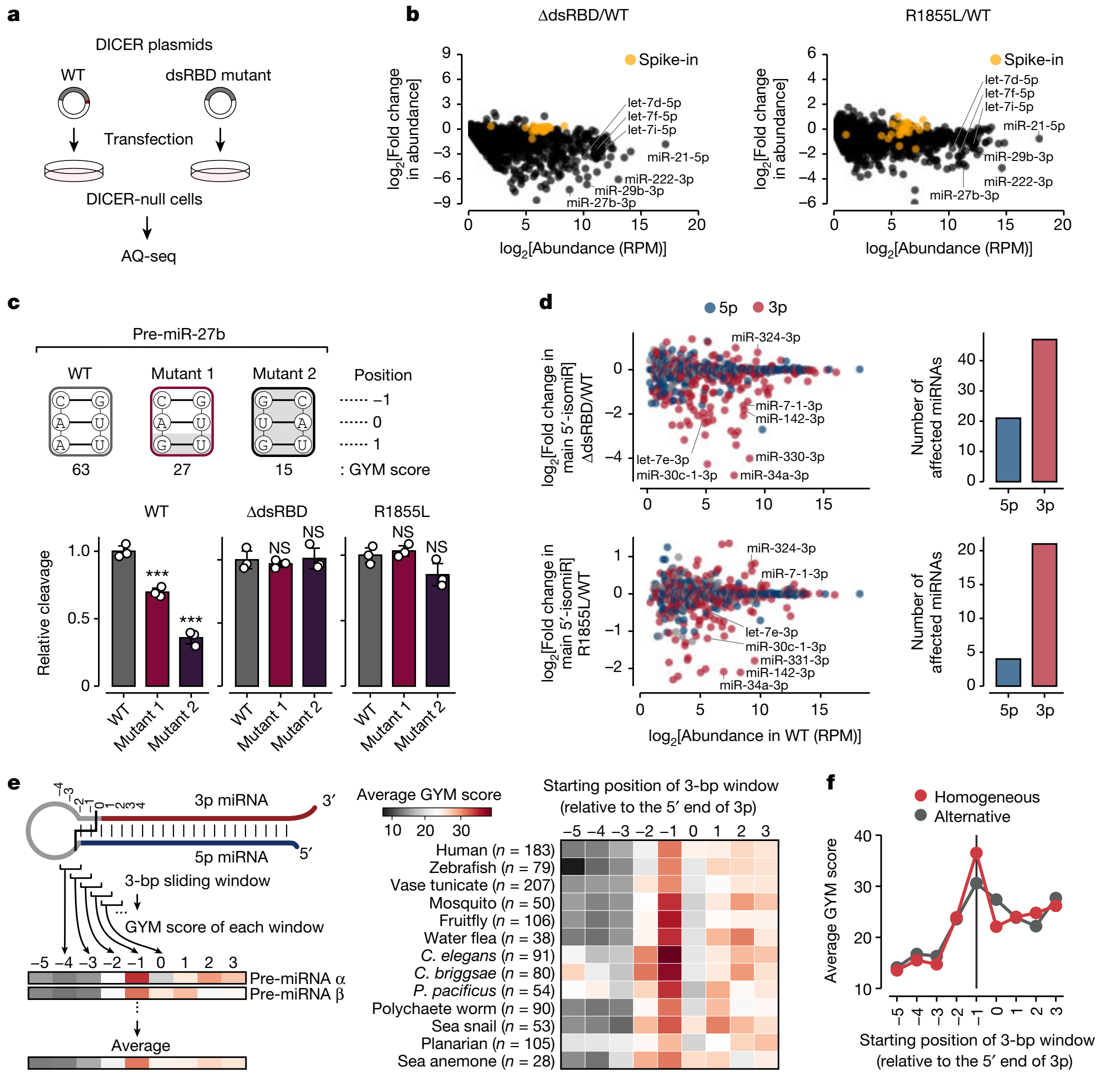


Fig. 3 | Endogenous miRNAs are regulated by the dsRBD in a GYM-motif-dependent fashion. a, Schematic outline of the rescue experiment ($n = 2$, biological replicates). **b**, Comparison of miRNA abundance. Spike-ins were used for normalization. RPM, reads per million. **c**, In vitro processing of pre-miR-27b variants by either DICER WT or dsRBD mutants. RNA substrates, radiolabelled at their 5′ end, vary in their sequences at positions −1 to 1, having different GYM scores (top). Relative cleavage was calculated by quantifying the band intensity (1 − uncleaved/input) (bottom). Bars indicate mean ± s.d. ($n = 3$, independent experiments). ***$P < 0.001$ by two-sided Student's t-test compared to the WT substrate. **d**, Comparison of cleavage accuracy (left). For a given miRNA, the most abundant 5′-isomiR was identified in the WT sample. Then the fold change of its proportions was quantified to estimate cleavage accuracy. Grey, unannotated strand. Bar graphs show the number of miRNAs whose main 5′-isomiR was significantly affected by the mutation ($P < 0.01$ by two-sided Student's t-test) (right). **e**, Schematic outline of conservation analysis of GYM motifs on pre-miRNAs (left). Distribution of the GYM motif at the indicated positions in natural pre-miRNAs across species (right). *C. elegans*, *Caenorhabditis elegans*; *C. briggsae*, *Caenorhabditis briggsae*; *P. pacificus*, *Pristionchus pacificus*. **f**, The association between the GYM score and alternative processing. miRNAs registered in miRGeneDB were included in this analysis. For a given miRNA, if the proportion of the most abundant 5′-isomiR is over 80%, it was classified into the homogeneous processing group ($n = 203$). Otherwise, it was classified into the alternative processing group ($n = 76$).

cleavage site choice, compare Fig. 2b with Fig. 2h and Extended Data Fig. 6f). Taken together, these findings indicate that DICER recognizes the mismatched M mainly using its highly conserved R1855 and E1859 residues of the dsRBD to ensure efficient and precise processing.

Role of GYM motif in miRNA biogenesis

We next asked whether the GYM motif is biologically relevant in the context of endogenous miRNA maturation. The wild-type or mutant DICER proteins were ectopically expressed in *DICER*-knockout HCT116 or HEK293T cells (Fig. 3a). Small RNAs were sequenced by the AQ-seq protocol for bias-minimized quantification that allows reliable comparison between miRNA isoforms (isomiRs)[25] (Extended Data Fig. 7).

The miRNA abundance was reduced when the dsRBD was deleted or when the R1855 and E1859 residues were mutated (Fig. 3b and Extended Data Fig. 7a,c). Some conserved and abundant miRNAs were strongly affected, prompting us to investigate whether these miRNAs are produced in a GYM-dependent manner. In vitro processing assays showed that alterations in the dsRBD reduced the DICER activity on pre-miR-27b, pre-miR-21, pre-let-7d, pre-let-7f-1 and pre-let-7i, thus requiring a longer incubation time (Fig. 3c and Extended Data Fig. 8). Weakening variations in the GYM motif of the pre-miRNAs decreased processing efficiency, and this GYM motif effect was diminished when the dsRBD was mutated. Collectively, our data demonstrate that the interaction between the GYM motif and the DICER dsRBD promotes miRNA processing both in cells and in vitro.

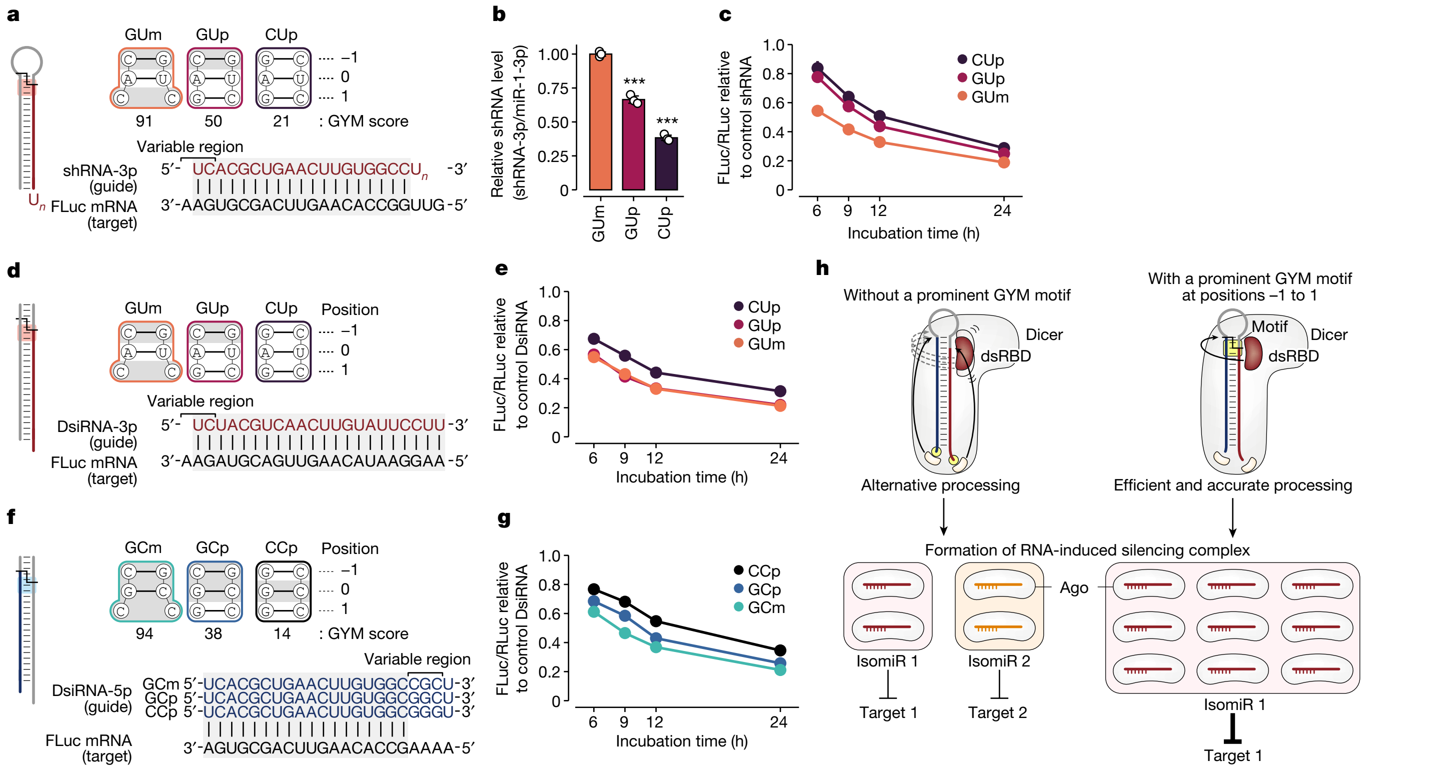


Fig. 4 | The GYM motif enhances gene silencing potency. a, Design of 3p-dominant shRNAs with optimal (GUm), suboptimal (CUp) and poor (CUp) GYM motifs. The sequence variation does not affect the mature siRNA sequence. shRNA was encoded in a plasmid that also encoded pri-miR-1-1 under an independent promoter. 'U_n' indicates a varying number of uridines that originate from the RNA polymerase III termination signal. FLuc, firefly luciferase. **b**, shRNA levels at 12 h measured by quantitative real-time PCR. The TaqMan probe was designed to target the mature shRNA that is the same among the variants. Bars indicate mean ± s.d. (n = 3, biological replicates). ***P < 0.001 by two-sided Student's t-test compared to GUm. **c**, Luciferase assay. FLuc signals were normalized to *Renilla* luciferase (RLuc) signals. A control shRNA that does not target FLuc was used to measure gene silencing activities. Points and bars indicate mean ± s.d. (n = 3, biological replicates). **d**, Design of 3p-dominant DsiRNAs with varying GYM motifs. **e**, Luciferase assay. Points and bars indicate mean ± s.d. (n = 3, biological replicates). **f**, Design of 5p-dominant DsiRNAs with varying GYM motifs. **g**, Luciferase assay. Points and bars indicate mean ± s.d. (n = 3, biological replicates). **h**, A proposed model of Dicer processing.

We also detected substantial alterations in the cleavage sites. The most notable example was miR-324-3p, which is known to be dependent on the dsRBD of DICER in vitro[27] (Extended Data Fig. 9a). In addition, we found many miRNAs whose processing was affected by the dsRBD alterations (Fig. 3d and Extended Data Fig. 9b). To infer the cleavage sites for DROSHA and DICER, we examined the 5′ ends of 5p miRNAs (Fig. 3d, blue dots) and 3p miRNAs (Fig. 3d, red dots), respectively. For quantification, we measured the proportion of the main 5′-isomiR (that is, isomiRs with the same 5′ end). The DICER dsRBD alterations affected mainly the DICER cleavage sites, leading to seed alterations and/or strand switches of many miRNAs[27] (Fig. 3d and Extended Data Fig. 9). We observed similar changes in HEK293T cells (Extended Data Fig. 7d). For validation, we carried out in vitro assays with pre-miR-34a and pre-let-7e and found that their cleavage patterns markedly changed when the dsRBD was deleted, consistently with the sequencing data (Fig. 3d and Extended Data Fig. 10a–d). The GYM motif alterations changed processing sites in these pre-miRNAs, confirming the importance of the GYM motif in cleavage site choice. Notably, these GYM variants cannot be distinguished by the ΔdsRBD mutant, indicating that the dsRBD–GYM interaction is critical for the cleavage site decision in human pre-miRNA processing. It is noteworthy that the GYM motif recognition seems to have differential effects on different miRNAs, either increasing (for example, pre-miR-7-1 and pre-miR-30c) or decreasing (for example, pre-miR-34a and pre-miR-324) the cleavage homogeneity (Extended Data Figs. 9a,b and 10a,b), which may be because the GYM motif cooperates or competes with the end counting mechanisms depending on its relative position.

Evolutionary implications

If the GYM motif plays an important and general role in miRNA maturation, one would expect that GYM motifs are pronounced at position −1 to 1 across multiple pre-miRNAs and various species. Indeed, in human pre-miRNAs, the GYM scores are higher at positions −1 to 1 compared to those at neighbouring positions (Fig. 3e, human). Notably, such position-specific enrichment was observed in all metazoan species examined (that is, from sea anemone to human). Together with our biochemical data showing that fly Dcr-1 is facilitated by the GYM motif in vitro (Fig. 2c,d), this positional conservation pattern suggests that the GYM motif is deeply rooted in eumetazoan miRNA evolution.

Notably, homogeneously processed miRNAs show stronger positional enrichment at a single site than alternatively processed miRNAs do (Fig. 3f). This observation suggests that the GYM motifs that are 'prominent' relative to the neighbouring positions facilitate the production of single miRNA species, whereas 'not-so-prominent' GYM motifs allow alternative processing, yielding multiple isomiRs and diversifying miRNA functions, which possibly explains why most, but not all, human pre-miRNAs evolved to have prominent GYM motifs (Extended Data Fig. 10e).

To experimentally test this hypothesis, we selected pre-miR-9, which is alternatively processed by DICER in cells to produce main 22-nt and minor 21-nt isoforms (Extended Data Fig. 10f, blue and green arrowheads, respectively). Notably, pre-miR-9 lacks a prominent GYM motif, with comparable GYM scores between two 3-bp sliding windows corresponding to the two isomiRs. We introduced a weaker GYM variant

specifically to the lower window, without altering the GYM motif in the upper window, which in effect made the upper motif more 'prominent' without increasing its score. In this situation, pre-miR-9 was no longer alternatively processed, indicating that the comparable GYM motifs in two consecutive windows allowed alternative processing of pre-miR-9 (Extended Data Fig. 10g). Thus, depending on their relative strength and position, GYM motifs can be used to facilitate either homogeneous or alternative processing of pre-miRNAs.

The GYM motif improves RNA interference

Having identified the GYM motif as a strong determinant of DICER-mediated processing, we questioned its effect on RNA interference by short hairpin RNA (shRNA) and Dicer-substrate siRNA (DsiRNA) that are dependent on cellular Dicer to yield siRNAs. shRNA mimics pre-miRNA but contains a 5′ triphosphate group and a 3′ overhang with 3–5 uridines originating from the RNA polymerase III termination signal[33]. This non-optimal terminal structure impairs DICER processing, posing a drawback of shRNA-mediated interference. To test whether the GYM motif can override the end-dependency, we constructed shRNAs with GYM motifs of variable scores (Fig. 4a). The shRNA with a high-scoring GYM motif yielded the highest level of siRNA products (Fig. 4b) and the most potent gene silencing activity (Fig. 4c).

We also evaluated the GYM motif in the context of DsiRNA—a synthetic 27-nt siRNA duplex designed to be processed by DICER to yield 22-nt siRNA[34]. DsiRNAs exhibit more potent activity than synthetic 21-nt siRNA duplexes that bypass Dicer processing, suggesting a coupling between Dicer processing and Ago loading[35,36]. DsiRNAs can be designed to place the antisense guide sequence in either strand. Strong GYM motifs increased the knockdown efficiencies, regardless of whether the antisense is placed in the 5p or 3p strand (Fig. 4d–g). Taking these findings together, the identified GYM motif promotes both shRNA- and DsiRNA-mediated interference, allowing effective knockdown.

Discussion

Here we used massively parallel assays to identify a position-dependent sequence determinant for DICER processing, namely the GYM motif. The GYM score offers a way to quantitatively define optimal substrates of DICER. The GYM-mediated mechanism is conserved across metazoan species. It remains to be investigated whether a similar mechanism is at work in plants and fungi and if this also applies to siRNA-specific DICER homologues such as fly Dicer-2. Moreover, it will be intriguing to interrogate the remaining regions of pre-miRNAs—the loop and the lower part of the stem—to examine additional sequence determinants.

Central to the GYM motif-mediated mechanism is the DICER dsRBD that recognizes the element. This contrasts with the dsRBDs of the DICER partner protein, TRBP, that do not show sequence specificity[37]. The DICER dsRBD was known to enhance cleavage efficiency[7,38]. Our study reveals how the dsRBD achieves its function by recognizing the substrate RNA in not only sequence-independent but also sequence-dependent manners. Furthermore, this study unveils the function of the dsRBD in the cleavage site decision, and shows that it can override the function of the other domains. In light of our finding, it is noteworthy that dsRBDs of bacterial RNase III (class I RNase III) and human DROSHA (class II RNase III) also contribute to substrate specificity by recognizing the sequence and structural features near the cleavage sites although the positions and sequences of the motifs are different[29,39,40]. Therefore, together with the previous reports, the current study expands our understanding of dsRBDs by illustrating their high amenability conferring specificity to a wide array of sequence and structural compositions in dsRNA.

The GYM motif is a decisive factor specifying the cleavage site, which can override the 5′ and 3′ end counting rules when in conflict. However, when consistently positioned relative to the termini, the GYM motif can cooperate with the termini to ensure accurate and homogeneous processing. Taking our results together, we propose a model for the role of the GYM motif in small RNA biogenesis and its effects on RNA silencing (Fig. 4h). For a pre-miRNA with a not-so-prominent GYM motif, the dsRBD of Dicer does not form a tight interaction at a specific position (Fig. 4h, top left) so the processing accuracy is compromised, generating multiple isoforms. This mechanism can offer a useful basis for regulation through alternative processing that can expand and/or alter the target repertoire by changing the seed sequence and sometimes even switching the functional strand (a phenomenon known as arm switching) as seen with miR-9 and miR-324 (ref. [27]). In situations in which there is a prominent GYM motif, the dsRBD latches onto the GYM motif to accommodate the dsRNA in a fixed position (Fig. 4h, top right). In this case, the impact of the end counting rule can be negated. Thus, this study amends the previous notion that the Dicer cleavage site is passively determined by Drosha, which generates pre-miRNA termini. Accordingly, the GYM motif can compensate for the imprecise (alternative) processing of pri-miRNAs by Drosha[25,26,28,40–43] or negate the effect of promiscuous end modifications (such as 3′ tailing and trimming)—which would otherwise affect Dicer cleavage sites[25,26,44]. This allows Dicer to precisely process the substrate to produce miRNAs and siRNAs with high accuracy and efficiency, which consequently contributes to the fidelity and potency of RNA silencing.

In light of our model for the critical role of the GYM motif at the catalytic step, we set out to determine the long-sought-after structure of human DICER in a dicing state using a pre-miRNA with a high-scoring GYM motif[45]. The cryogenic electron microscopy structure at a resolution of 3.0 Å revealed that the dsRBD indeed forms sequence-specific interactions with the GYM motif through R1855 (ref. [45]), correlating our in vitro and cellular findings.

Together, the findings of our study reveal an integral and conserved mechanism of substrate recognition by DICER and provide a framework to understand how DICER produces small RNAs for biological and therapeutic regulation. Moreover, our finding of a cancer-associated substitution (R1855L) in DICER that disrupts the recognition of the GYM motif offers insights into the molecular impact of this substitution.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-05722-4.

1. Elbashir, S. M., Lendeckel, W. & Tuschl, T. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.* **15**, 188–200 (2001).
2. Hutvagner, G. et al. A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science* **293**, 834–838 (2001).
3. Macrae, I. J. et al. Structural basis for double-stranded RNA processing by Dicer. *Science* **311**, 195–198 (2006).
4. Macrae, I. J., Li, F., Zhou, K., Cande, W. Z. & Doudna, J. A. Structure of Dicer and mechanistic implications for RNAi. *Cold Spring Harb. Symp. Quant. Biol.* **71**, 73–80 (2006).
5. Park, J. E. et al. Dicer recognizes the 5′ end of RNA for efficient and accurate processing. *Nature* **475**, 201–205 (2011).
6. Zhang, H., Kolb, F. A., Jaskiewicz, L., Westhof, E. & Filipowicz, W. Single processing center models for human Dicer and bacterial RNase III. *Cell* **118**, 57–68 (2004).
7. Zhang, H., Kolb, F. A., Brondani, V., Billy, E. & Filipowicz, W. Human Dicer preferentially cleaves dsRNAs at their termini without a requirement for ATP. *EMBO J.* **21**, 58755885 (2002).
8. Gu, S. et al. The loop position of shRNAs and pre-miRNAs is critical for the accuracy of Dicer processing in vivo. *Cell* **151**, 900–911 (2012).
9. Tsutsumi, A., Kawamata, T., Izumi, N., Seitz, H. & Tomari, Y. Recognition of the pre-miRNA structure by *Drosophila* Dicer-1. *Nat. Struct. Mol. Biol.* **18**, 1153–1158 (2011).
10. Feng, Y., Zhang, X., Graves, P. & Zeng, Y. A comprehensive analysis of precursor microRNA cleavage by human Dicer. *RNA* **18**, 2083–2092 (2012).
11. Liu, Z., Wang, J., Li, G. & Wang, H. W. Structure of precursor microRNA's terminal loop regulates human Dicer's dicing activity by switching DExH/D domain. *Protein Cell* **6**, 185–193 (2015).

12. Lee, Y., Jeon, K., Lee, J. T., Kim, S. & Kim, V. N. MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J.* **21**, 4663–4670 (2002).
13. Lee, Y. et al. The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425**, 415–419 (2003).
14. Denli, A. M., Tops, B. B., Plasterk, R. H., Ketting, R. F. & Hannon, G. J. Processing of primary microRNAs by the Microprocessor complex. *Nature* **432**, 231–235 (2004).
15. Gregory, R. I. et al. The Microprocessor complex mediates the genesis of microRNAs. *Nature* **432**, 235–240 (2004).
16. Han, J. et al. The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev.* **18**, 3016–3027 (2004).
17. Khvorova, A., Reynolds, A. & Jayasena, S. D. Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115**, 209–216 (2003).
18. Schwarz, D. S. et al. Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115**, 199–208 (2003).
19. Liu, J. et al. Argonaute2 is the catalytic engine of mammalian RNAi. *Science* **305**, 1437–1441 (2004).
20. Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. Prediction of mammalian microRNA targets. *Cell* **115**, 787–798 (2003).
21. Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215–233 (2009).
22. Tian, Y. et al. A phosphate-binding pocket within the platform-PAZ-connector helix cassette of human Dicer. *Mol. Cell* **53**, 606–616 (2014).
23. Liu, Z. et al. Cryo-EM structure of human Dicer and its complexes with a pre-miRNA substrate. *Cell* **173**, 1191–1203 (2018).
24. Heo, I. et al. Mono-uridylation of pre-microRNA as a key step in the biogenesis of group II let-7 microRNAs. *Cell* **151**, 521–532 (2012).
25. Kim, H. et al. Bias-minimized quantification of microRNA reveals widespread alternative processing and 3′ end modification. *Nucleic Acids Res.* **47**, 2630–2640 (2019).
26. Chiang, H. R. et al. Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev.* **24**, 992–1009 (2010).
27. Kim, H. et al. A mechanism for microRNA arm switching regulated by uridylation. *Mol. Cell* **78**, 1224–1236 (2020).
28. Kim K. et al. A quantitative map of human primary microRNA processing sites. *Mol. Cell* **81**, P3422–3439.E11 (2021).
29. Fang, W. & Bartel, D. P. The menu of features that define primary microRNAs and enable de novo design of microRNA genes. *Mol. Cell* **60**, 131–145 (2015).
30. Han, C. et al. The RNA-binding protein DDX1 promotes primary microRNA maturation and inhibits ovarian tumor progression. *Cell Rep.* **8**, 1447–1460 (2014).
31. Cerami, E. et al. The cBio Cancer Genomics Portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
32. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1 (2013).
33. Gao, Z., Herrera-Carrillo, E. & Berkhout, B. Delineation of the exact transcription termination signal for type 3 polymerase III. *Mol. Ther. Nucleic Acids* **10**, 36–44 (2018).
34. Amarzguioui, M. et al. Rational design and in vitro and in vivo delivery of Dicer substrate siRNA. *Nat. Protoc.* **1**, 508–517 (2006).
35. Kim, D. H. et al. Synthetic dsRNA Dicer substrates enhance RNAi potency and efficacy. *Nat. Biotechnol.* **23**, 222–226 (2005).
36. Snead, N. M. et al. Molecular basis for improved gene silencing by Dicer substrate interfering RNA compared with other siRNA variants. *Nucleic Acids Res.* **41**, 6209–6221 (2013).
37. Masliah, G. et al. Structural basis of siRNA recognition by TRBP double-stranded RNA binding domains. *EMBO J.* **37**, e97089 (2018).
38. Ma, E., Zhou, K., Kidwell, M. A. & Doudna, J. A. Coordinated activities of human Dicer domains in regulatory RNA processing. *J. Mol. Biol.* **422**, 466–476 (2012).
39. Gan, J. et al. Structural insight into the mechanism of double-stranded RNA processing by ribonuclease III. *Cell* **124**, 355–366 (2006).
40. Kwon, S. C. et al. Molecular basis for the single-nucleotide precision of primary microRNA processing. *Mol. Cell* **73**, 505–518 (2019).
41. Kim, B., Jeong, K. & Kim, V. N. Genome-wide mapping of DROSHA cleavage sites on primary microRNAs and noncanonical substrates. *Mol. Cell* **66**, 258–269 (2017).
42. Bofill-De Ros, X. et al. Structural differences between pri-miRNA paralogs promote alternative Drosha cleavage and expand target repertoires. *Cell Rep.* **26**, 447–459 (2019).
43. Wu, H., Ye, C., Ramirez, D. & Manjunath, N. Alternative processing of primary microRNA transcripts by Drosha generates 5′ end variation of mature microRNA. *PLoS ONE* **4**, e7566 (2009).
44. Burroughs, A. M. et al. A comprehensive survey of 3′ animal miRNA modification events and a possible role for 3′ adenylation in modulating miRNA targeting effectiveness. *Genome Res.* **20**, 1398–1410 (2010).
45. Lee, Y.-Y., Lee, H., Kim, H., Kim, V. N. & Roh S.-H. Structure of the human DICER–pre-miRNA complex in a dicing state. *Nature* https://doi.org/10.1038/s41586-023-05723-3 (2023).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2023

Methods

Plasmid construction

To construct plasmids for expressing human DICER, the coding sequence of the *DICER* reference (RefSeq NM_030621) was amplified and subcloned into the pX vector (for protein purification) containing an N-terminal His_{10}-eYFP-SUMOstar-Strep sequence, or the pCK vector (for ectopic expression in *DICER*-knockout cells) with a PGK promoter substituted for the CMV promoter. Other DICER constructs, including DICER(ΔdsRBD) (amino acids 1–1,848), were generated through site-directed mutagenesis. For the plasmid expressing human TRBP, the coding sequence of the human *TRBP* reference (RefSeq NM_134323) was amplified and subcloned into the same pX vector. For the plasmid expressing *Drosophila* Dcr-1, the coding sequence of the *Drosophila Dcr-1* reference (RefSeq NM_079729.3) was amplified from *Drosophila* cDNA and subcloned into the same pX vector. To construct plasmids for dual reporter luciferase assays, synthetic DNA oligonucleotides containing the pri-miRNA sequences or the target sites of shRNAs and DsiRNAs were inserted into the pmirGLO vector (Promega), downstream of the FLuc gene. To construct plasmids for shRNA expression, synthetic DNA oligonucleotides containing the shRNA sequences and pri-miR-1-1 sequence were inserted into a vector, downstream of U6 and UbC promoters, respectively.

Protein purification

N-terminal His_{10}-eYFP-SUMOstar-Strep-tagged DICER expression plasmids were transiently transfected into HEK293E cells, derived from human embryonic kidneys, that were grown in suspension culture (Dulbecco's modified Eagle's medium (DMEM) supplemented with 5% fetal bovine serum) in a 8% CO_2 humidified shaking incubator at 37 °C. At a cell density of about 7×10^6 cells per millilitre, each plasmid was transfected at a final concentration of 0.3 μg ml^{-1} with linear polyethylenimine and dimethylsulfoxide sequentially added at a final concentration of 3 μg ml^{-1} and 1%, respectively. Transfected cells were incubated at 33 °C for 72 h.

All purification procedures were carried out at 4 °C. The cell pellet was resuspended in buffer containing 100 mM Tris-HCl (pH 8.0), 150 mM NaCl, 10% glycerol and 2 mM β-mercaptoethanol (BME), supplemented with 0.1 mM phenylmethylsulfonyl fluoride (PMSF), EDTA-free Pierce Protease Inhibitor Mini Tablets (Thermo Fisher Scientific), 20 μl ml^{-1} micrococcal nuclease and 5 mM $CaCl_2$. Cells were lysed by sonication and centrifuged at 35,000*g* for 1 h. The supernatant was applied to Ni-NTA Superflow resin (Qiagen) equilibrated with buffer containing 100 mM Tris-HCl (pH 8.0), 150 mM NaCl, 10% glycerol, 2 mM BME and 0.1 mM PMSF. The resin was washed with 5 column volumes of the equilibrium buffer supplemented with 40 mM imidazole. The bound proteins were eluted with the equilibrium buffer supplemented with 200 mM imidazole. For proteolytic cleavage of the N-terminal His_{10}-eYFP-SUMOstar tag, the sample was treated with SUMOstar protease (LifeSensors) at 4 °C for 1 h. The sample was then applied to Strep-Tactin Superflow (IBA Lifesciences) resin equilibrated with buffer containing 100 mM Tris-HCl (pH 8.0), 150 mM NaCl, 2 mM BME and 0.1 mM PMSF. The resin was washed sequentially with 5 column volumes of the equilibrium buffer containing 2 mM EDTA and no EDTA. The bound protein was eluted with buffer containing 100 mM Tris-HCl (pH 8.0), 150 mM NaCl, 2 mM BME and 50 mM biotin. The eluate was applied to Ni-NTA Superflow resin (Qiagen) to remove uncleaved fusion protein. The unbound fraction was concentrated to about 20 μM using an Amicon Ultra-15 Centrifugal Filter Unit (100 kDa cutoff) and loaded onto Superose 6 Increase 5/150 GL (GE Healthcare) equilibrated with 50 mM Tris (pH 8.0), 100 mM NaCl and 0.5 mM TCEP. The fractions containing the protein were pooled, concentrated to about 5 μM, snap frozen in liquid nitrogen, and stored at −80 °C until use. All human DICER constructs and *Drosophila* Dcr-1 were purified using the same purification method. Protein concentrations were quantified by ultraviolet absorbance at 280 nm (NanoDrop), using a molecular extinction coefficient calculated for each protein.

Massively parallel assay to identify DICER processing determinants

To predict the structure of dsRNA-bound human DICER, the cryogenic electron microscopy structure of human DICER in the cleavage-incompetent state was superimposed with the crystal structure of *A. aeolicus* RNase III bound to dsRNA[39] using PyMol. On the basis of the structural model, the dsRNA region likely to be recognized by the dsRBD was predicted (Extended Data Fig. 2a). Synthetic pre-miRNAs (Integrated DNA Technologies), pre-let-7a-1 and pre-miR-374b, containing random bases within the specified window, were denatured at 80 °C for 5 min and annealed by slowly cooling the temperature down to 4 °C at a −1 °C min^{-1} rate in a buffer containing 20 mM Tris (pH 7.5), 80 mM NaCl and 1 mM EDTA. Massively parallel assays were carried out by mixing randomized pre-miRNA pool with purified DICER at final concentrations of 0.1 μM and 0.2 μM, respectively, in a reaction buffer containing 50 mM Tris (pH 7.5), 100 mM NaCl, 0.5 mM TCEP and 2 mM $MgCl_2$. The reaction was carried out at 37 °C for predetermined time periods to achieve approximately 10, 20 or 30% cleavage of the pre-miRNA pool. The reaction was stopped by adding an equivalent volume of 2X RNA Loading Dye (NEB) supplemented with additional 20 mM EDTA. Along with the input RNA, the mixture was resolved on a 15% urea–polyacrylamide gel. The input RNA and the uncleaved pre-miRNA pool were gel-purified. The uncleaved pre-miRNAs were subsequently subjected to AQ-seq as previously described[25], with some modifications as follows. After 3′ adapter ligation, the ligated RNAs were purified on a 10% urea–polyacrylamide gel along with Century-Plus RNA Markers (Thermo Fisher) as a size marker. After reverse transcription, cDNAs were amplified with NEBNext Ultra II Q5 Master Mix (NEB). The libraries were sequenced on the NovaSeq 6000 platform.

Sequencing reads were first pre-processed by clipping the 3′ adapter using cutadapt[46]. Next, 4-nt degenerate sequences were further removed with FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) and short, low-quality and artefactual reads were filtered out using FASTX-Toolkit. Only reads without mutations outside the randomized region were selected for subsequent analysis. For a given pre-miRNA variant, the cleavage score was calculated by dividing its proportion (with a pseudocount 1) in the input by its proportion in the uncleaved substrate sample. The sequences of pre-miRNAs are listed in Supplementary Table 1. The results are summarized in Supplementary Table 2.

In vitro processing of RNA substrates

Pre-miRNA substrates were prepared by ligating two single-stranded RNA (ssRNA) fragments (Integrated DNA Technologies) as previously described[47]. For preparation of dsRNA substrates, ssRNAs (Bioneer) of the 5′-arm were 5′-end-radiolabelled with [γ-^{32}P]ATP by T4 polynucleotide kinase (Takara) and purified using Oligo Clean & Concentrator (Zymo Research) according to manufacturer's instructions. Next, they were annealed to the complementary synthetic ssRNAs of the 3′-arm in a buffer containing 20 mM Tris (pH 7.5), 80 mM NaCl and 1 mM EDTA, by heating at 80 °C for 5 min and slowly cooling the temperature down to 4 °C at a −1 °C min^{-1} rate. In vitro processing assays were carried out by mixing the RNA substrates with the 0.5 μM purified proteins in a reaction buffer containing 50 mM Tris (pH 7.5), 100 mM NaCl, 0.5 mM TCEP, 2 mM $MgCl_2$ and 5 μg ml^{-1} yeast RNA. The reaction was carried out at 37 °C and stopped by adding an equivalent volume of 2X RNA Loading Dye (NEB) supplemented with 0.5 mg ml^{-1} proteinase K (Roche). The mixture was resolved on a 15% urea–polyacrylamide gel along with 5′-end-radiolabelled synthetic mature miRNAs and Decade Markers System (Ambion) as size markers and visualized by the phosphorimager, Typhoon FLA 7000 (GE Healthcare). The sequences of synthetic oligonucleotides are listed in Supplementary Table 1. Exact *P* values are presented in Supplementary Table 3.

DICER rescue experiment and data analysis

DICER-knockout HCT116 and HEK293T cells were maintained in DMEM and McCoy's 5A medium (WELGENE), respectively, supplemented with 10% fetal bovine serum (WELGENE)[48,49]. Both cell lines were authenticated using the ATCC short tandem repeat profiling and confirmed to be mycoplasma-negative. *DICER*-knockout HCT116 and HEK293T cells were transiently transfected with DICER expression plasmids using FuGENE HD (Promega) or Lipofectamine 3000 (Thermo Fisher), respectively. Total RNAs were extracted using TRIzol (Thermo Fisher) 48 h post-transfection and subjected to AQ-seq[25] except that 100 μg of total RNAs was used for library construction, and that cDNAs were amplified with NEBNext Ultra II Q5 Master Mix (NEB). The libraries were sequenced on the NovaSeq 6000 platform.

Data processing was carried out as previously described[25]. Briefly, the 3′ adapter and 4-nt degenerate sequences were removed from the reads using cutadapt[46] and FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), respectively. Next, short, low-quality and artefactual reads were filtered out using FASTX-Toolkit. The output reads were first mapped to the spike-in sequences, and then unmapped reads were aligned to the human genome (hg38) using BWA[50]. Corresponding miRNA annotations were found with miRBase release 21 using the intersect tool in BEDTools[51,52]. For the analysis of miRNA abundance, read counts of miRNAs were normalized with the average of read counts of spike-ins. For the analysis of processing accuracy, we first identified the most abundant 5′-isomiR for a given mature miRNA in samples expressing WT DICER. Then the proportions of the 5′-isomiR were measured for all of the samples. Fold changes of the proportions between samples were calculated to examine processing accuracy defects. To reduce the influence of decay intermediates, we excluded from subsequent analysis miRNAs that do not have any 5′-isomiR accounting for more than 20% of the 5′-isomiRs with 5′ ends within positions −5 to 5, relative to the annotated one. The results are summarized in Supplementary Tables 4 and 5.

Analysis on the GYM motif of natural miRNAs

Representative animal miRNAs from diverse species have been used in the primary miRNA motif conservation analysis[29,53]. We adopted the curated lists[29,53] for our analysis and further curated them as follows: miRNA annotations and sequences were manually curated on the basis of miRBase release 22 (ref. [51]); pre-miRNA sequences were retrieved on the basis of their 5p and 3p annotations in miRBase release 22 and added to the lists; pre-miRNA sequences from miRGeneDB release 2.0 were added to the lists[54]; 3p sequences from miRBase release 22 and miRGeneDB release 2.0 were added to the lists; starting positions of 3p miRNAs were identified for each miRNA if 3p sequences were available, and if not, 3p sequences and starting positions were manually curated on the basis of its orthologues in close species.

Pre-miRNA secondary structures were obtained using Fold in RNAstructure version 6.3 with default settings[55]. On the basis of the predicted structure for a given miRNA, the 3-bp composition of each position relative to the starting position of 3p was identified (Fig. 3e). Next, its corresponding cleavage score was obtained by calculating average GYM scores of the 3 bp measured in the massively parallel assay of two pre-miRNA backbones with randomization of −1 to 1 carried out under the condition in which approximately 20% of substrates were cleaved. Pre-miRNA sequences were referenced to those of miRGeneDB and miRBase and to those listed in ref. [29], in order of priority. 3p positions were referenced to those of miRGeneDB, miRBase, and our manual curation, in order of priority. If there was a bulge or mismatch for a given position, the position was excluded. If there was no available 3p position, the miRNA was excluded. The GYM motifs in human miRNAs are summarized in Supplementary Table 6. The curated lists of representative animal miRNAs are available in Supplementary Table 7.

Dual luciferase reporter assay for gene silencing activity

For shRNA, HEK293T cells were co-transfected with pmirGLO vector containing a target site of shRNA-3p along with a plasmid encoding both shRNA and pri-miR-1-1 under the control of independent promoters using Lipofectamine 3000 reagent (Thermo Fisher). shRNAs were optimally designed following the published guidelines[56]. For DsiRNA, HEK293T cells were co-transfected with pmirGLO containing a target site of DsiRNA along with DsiRNA (Integrated DNA Technologies) using Lipofectamine 2000 reagent (Thermo Fisher). Cells were collected at different time points post-transfection and the luciferase signals were measured by the Dual Luciferase Reporter Assay System according to the manufacturer's instructions (Promega) on a Spark Multimode Microplate Reader (TECAN). The sequences of shRNAs and DsiRNAs are listed in Supplementary Table 1. Exact *P* values are presented in Supplementary Table 3.

Quantitative real-time PCR

HEK293T cells were transfected as described above. RNAs were isolated at different time points post-transfection using TRIzol (Thermo Fisher) and Direct-zol RNA Miniprep Kit (Zymo Research) according to the manufacturer's instructions. cDNAs were synthesized using the TaqMan miRNA Reverse Transcription Kit (Applied Biosystems) and subjected to quantitative real-time PCR with the TaqMan MicroRNA Assay (Applied Biosystems) on the StepOnePlus RealTime PCR System (Thermo Fisher). Given that miR-1-3p is expressed at low levels in HEK293T (about 8 RPM in the AQ-seq result)[25], miR-1-3p was used as an internal control. Exact *P* values are presented in Supplementary Table 3.

Statistics and reproducibility

All in vitro assays were repeated at least twice (Figs. 2a–e,g,h and 3c and Extended Data Figs. 1c, 2c, 5a–c,e–g, 6a,c–f, 8a–c and 10b,d,g).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The massively parallel assay data and rescue data were deposited to the GEO repository (accession numbers GSE202535 and GSE215866). Other structural models cited in this study for analysis (5ZAL and 2EZ6) are also accessible on PDB. The Cancer Genome Atlas data for the *DICER* gene was accessed at cBioPortal (https://www.cbioportal.org).

Code availability

Custom analysis codes are available at https://github.com/haedong-kim615/dicer_gym_motif.

46. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
47. Heo, I. et al. TUT4 in concert with Lin28 suppresses microRNA biogenesis through premicroRNA uridylation. *Cell* **138**, 696–708 (2009).
48. Kim, Y. K., Kim, B. & Kim, V. N. Re-evaluation of the roles of DROSHA, Export in 5, and DICER in microRNA biogenesis. *Proc. Natl Acad Sci. USA* **113**, E1881–E1889 (2016).
49. Bogerd, H. P., Whisnant, A. W., Kennedy, E. M., Flores, O. & Cullen, B. R. Derivation and characterization of Dicer- and microRNA-deficient human cells. *RNA* **20**, 923–937 (2014).
50. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
51. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42**, D68–D73 (2014).
52. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
53. Fromm, B. et al. MirGeneDB 2.0: the metazoan microRNA complement. *Nucleic Acids Res.* **48**, D1172 (2020).
54. Auyeung, V. C., Ulitsky, I., McGeary, S. E. & Bartel, D. P. Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. *Cell* **152**, 844–858 (2013).

55. Bellaousov, S., Reuter, J. S., Seetin, M. G. & Mathews, D. H. RNAstructure: web servers for RNA secondary structure prediction and analysis. *Nucleic Acids Res.* **41**, W471–W474 (2013).
56. Bofill-De Ros, X. & Gu, S. Guidelines for the optimal design of miRNA-based shRNAs. *Methods* **103**, 157–166 (2016).

Acknowledgements We thank J.-S. Woo for the mammalian cell transfection protocol; B. Cullen for HEK293T *DICER*-knockout cell lines; M. Lee for *Drosophila* cDNA; B. Um, H. Jang, K. Kim, M. Kim, S. Son and Y. Park for valuable discussions; and Y.-G. Choi, S.-M. Ji, J. Yang, D.-E. Choi, S. Bang and E. Kim for technical assistance. This research was supported by Institute for Basic Science funding from the Ministry of Science and ICT of Korea (IBS-R008-D1 to Y.-Y.L., H.K. and V.N.K.), BK21 research fellowships from the Ministry of Education of Korea (to Y.-Y.L. and H.K.) and a National Research Foundation of Korea grant funded by the Korean government (NRF-2018-Global PhD Fellowship Program to Y.-Y.L. and NRF-2015-Global PhD Fellowship Program to H.K.).

Author contributions Y.-Y.L. carried out structural modelling and protein purification. Y.-Y.L. and H.K. carried out biochemical and cell biological experiments. H.K. carried out bioinformatic analyses. Y.-Y.L., H.K. and V.N.K. designed the study and wrote the paper.

Competing interests Y.-Y.L., H.K. and V.N.K. are coinventors on pending patent application (KR 10-2022-0059227), submitted by Institute for Basic Science and Seoul National University, which covers the use of the GYM motif for RNA interference.

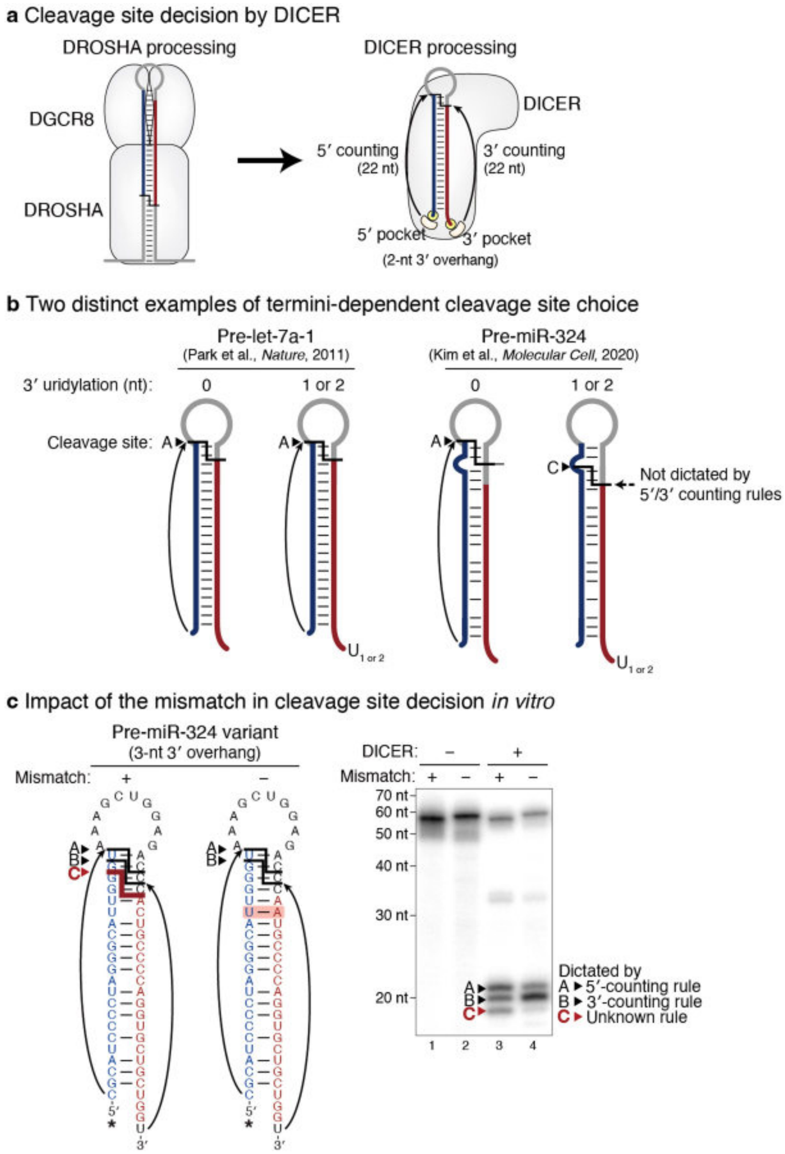
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-05722-4.

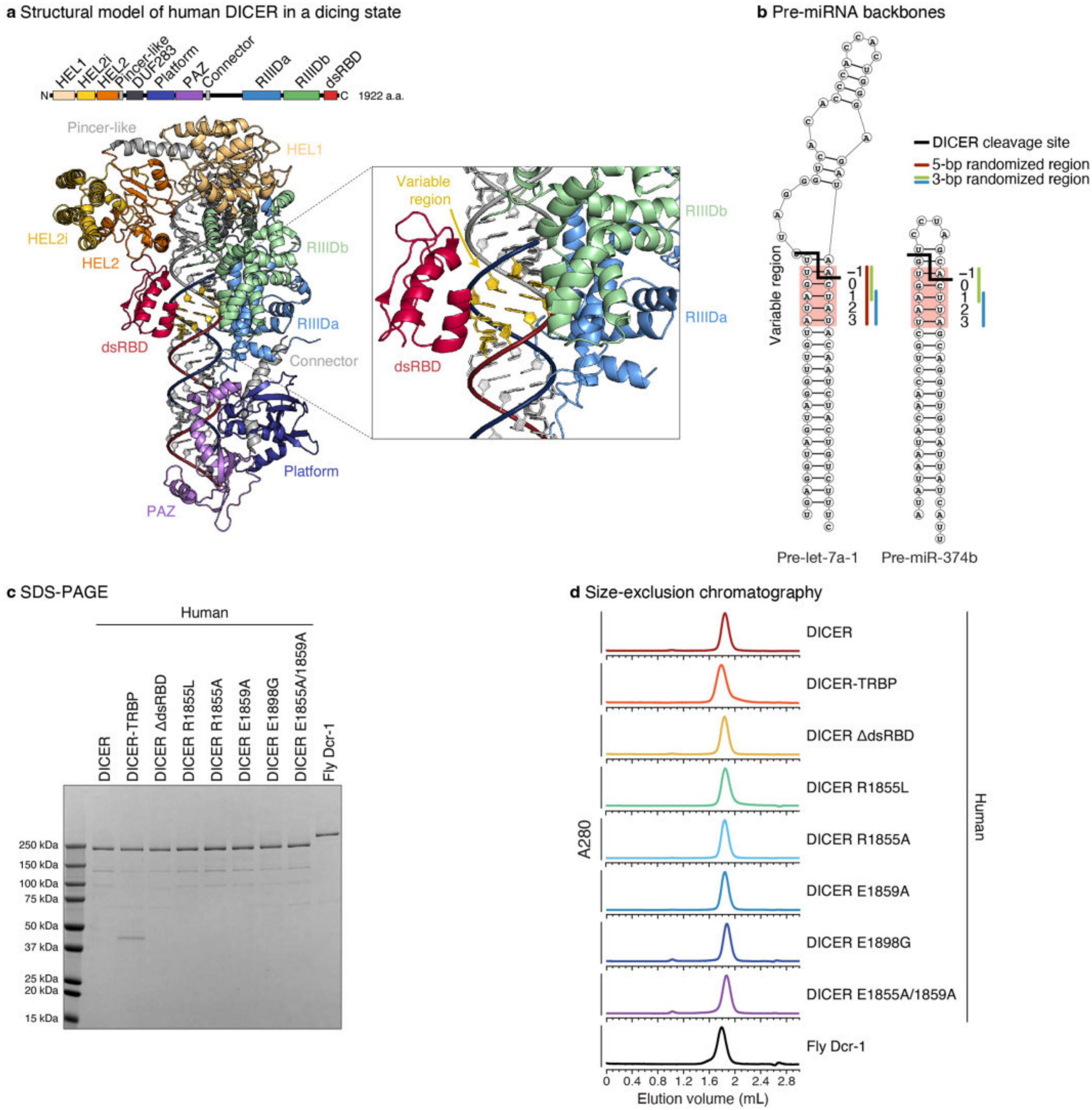
Correspondence and requests for materials should be addressed to V. Narry Kim.

Peer review information *Nature* thanks Haruhiko Siomi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

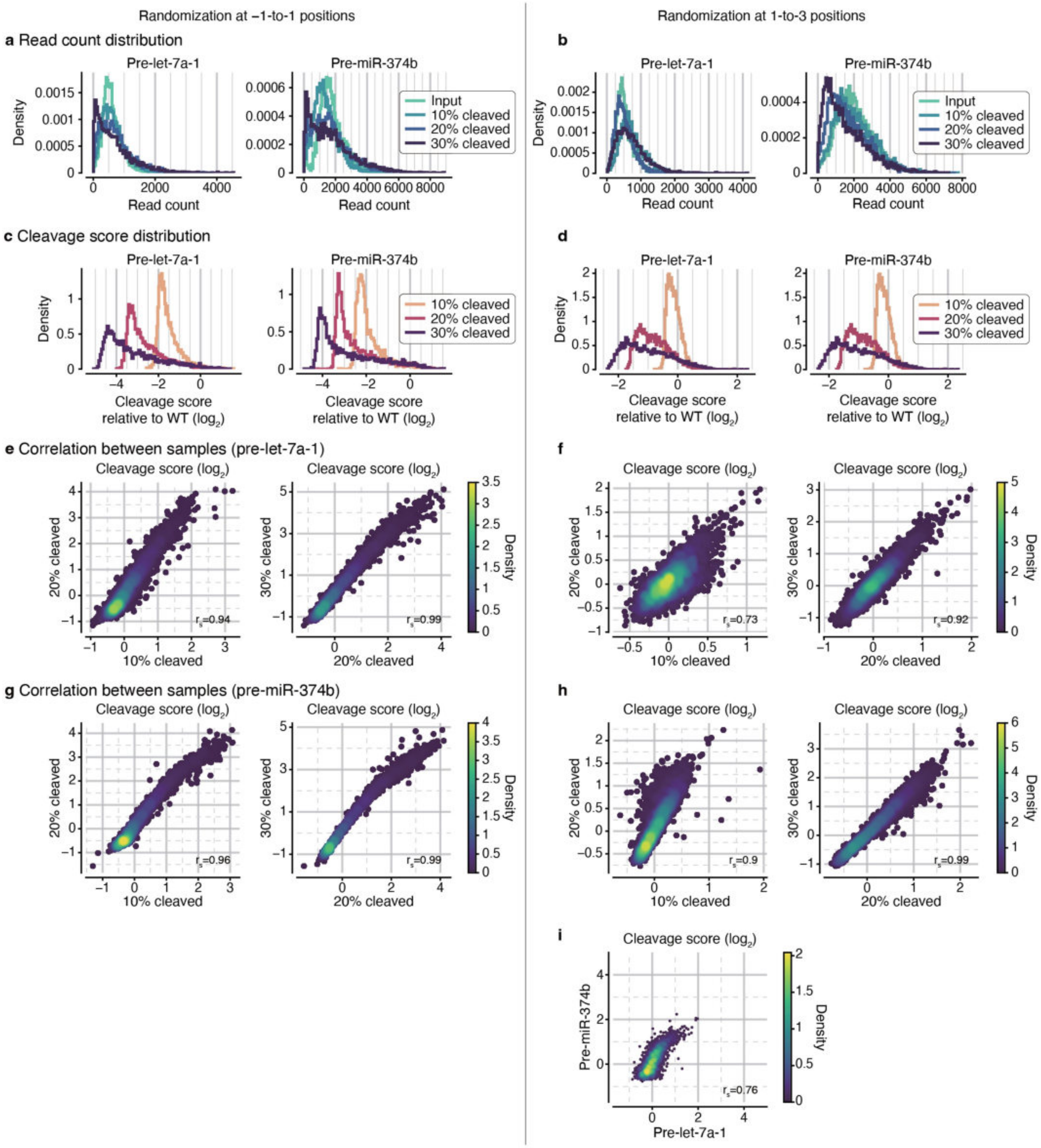
Reprints and permissions information is available at http://www.nature.com/reprints.



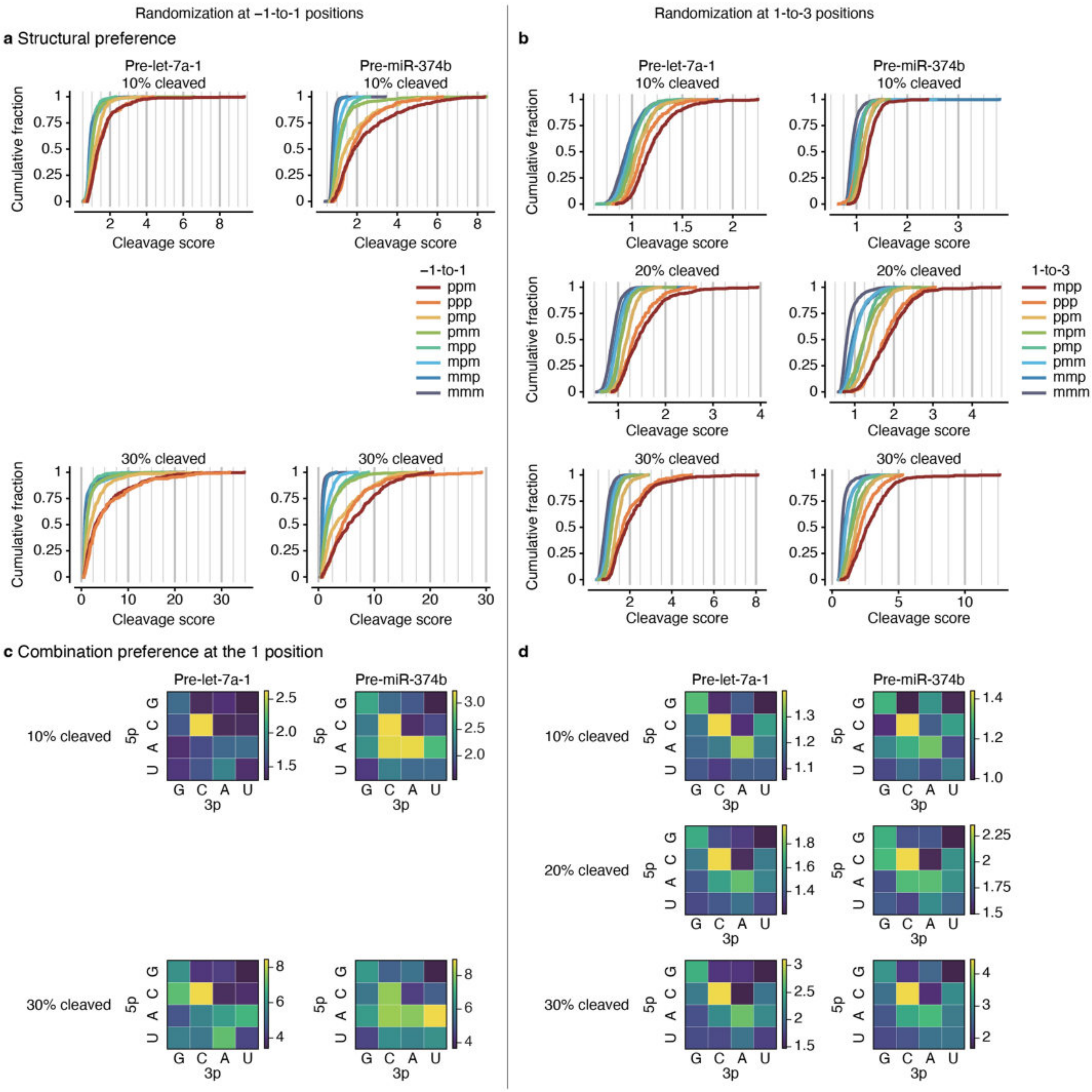
Extended Data Fig. 1 | A yet-unknown mechanism of DICER processing mediated by the upper stem region. a, Illustration of the mechanism of cleavage site choice by DICER. b, Cleavage site decision of pre-let-7a-1 and pre-miR-324. c, In vitro processing of a pre-miR-324 variant by DICER. "No-bulge pre-miR-324" was used for this assay to avoid the influence of the bulge[27]. A mismatch near the cleavage was replaced with a base-pair marked in pink. Cleavage sites and their corresponding products are marked with arrowheads. For gel source data, see Supplementary Fig. 1. *, radiolabeled 5′ phosphate.



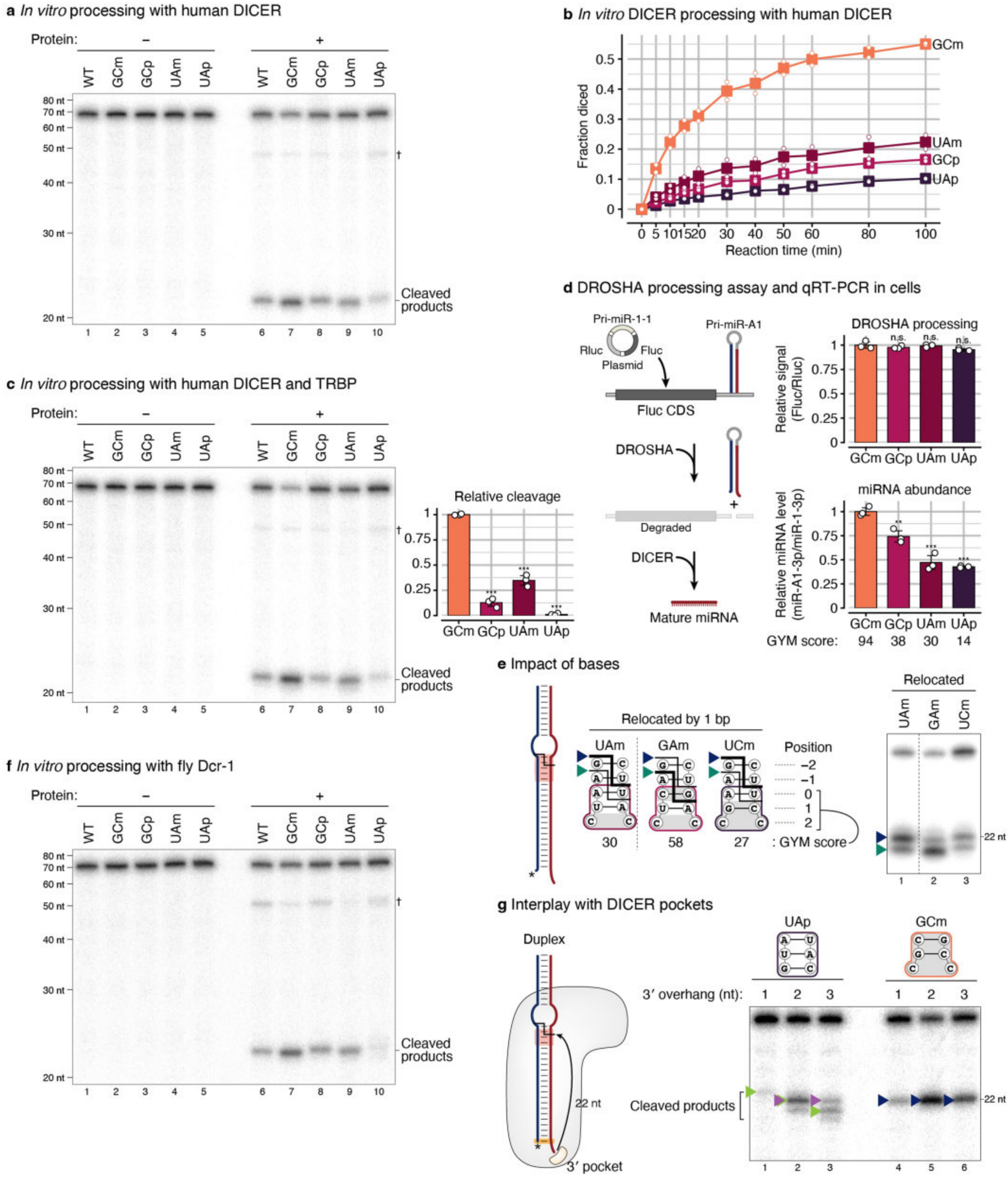
Extended Data Fig. 2 | Design of the massively parallel assay. a, A structural model of human DICER in a dicing state. The dsRNA was modeled into the cryo-EM structure of human DICER[23], based on the crystal structure of dsRNA-bound Aa RNase III[39]. DICER dsRBD was then superimposed with that of Aa RNase III to predict its position in a dicing state. b, Pre-miRNAs used in the massively parallel assay. The 5-bp and 3-bp windows (positions −1 to 3, −1 to 1, 1 to 3 relative to the starting position of 3p miRNA) were targeted for randomization based on the structural model. Secondary structures of pre-miRNAs were obtained using RNAstructure[52]. c, SDS-PAGE of purified proteins. For gel source data, see Supplementary Fig. 1. d, Size-exclusion chromatography of purified proteins.



Extended Data Fig. 3 | Massively parallel assays performed with substrates with −1-to-1 or 1-to-3 randomization. a–b, Distribution of read counts of variants. c–d, Distribution of cleavage scores of variants. e–h, Correlation of cleavage scores of variants between different conditions of varying reaction time. i, Distribution of the cleavage scores measured from the 2nd screening with 1-to-3 randomization.

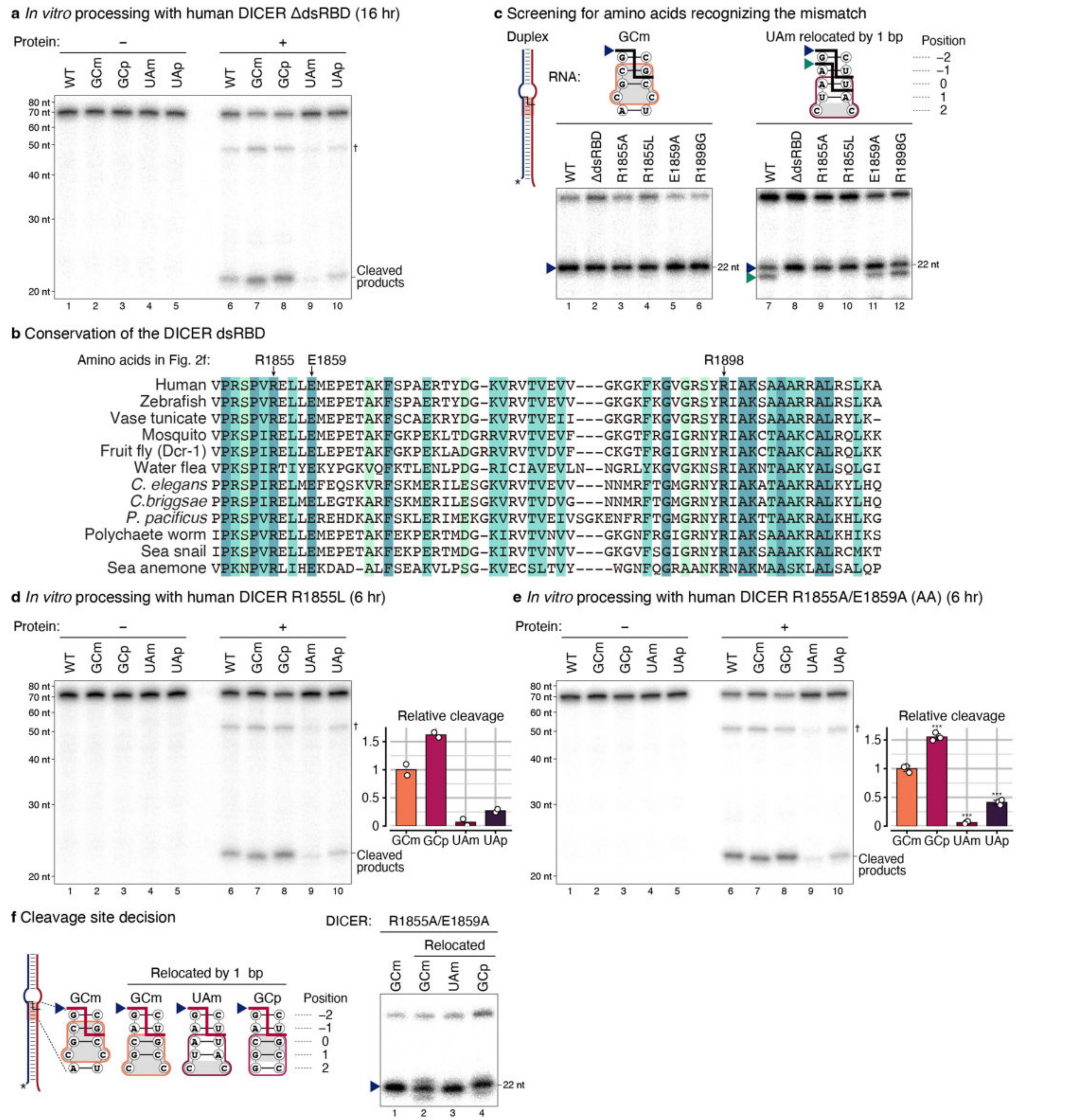


Extended Data Fig. 4 | Massively parallel assay reveals structural and sequence preferences at position 1. a–b, Structural impact on cleavage scores. G–U pair was considered as a mismatch only when it is in between mismatches. p, pair; m, mismatch. c–d, Impact of the base combinations at the 1 position on cleavage scores. Variants with base-pairs at all but position 1 were included in this analysis.

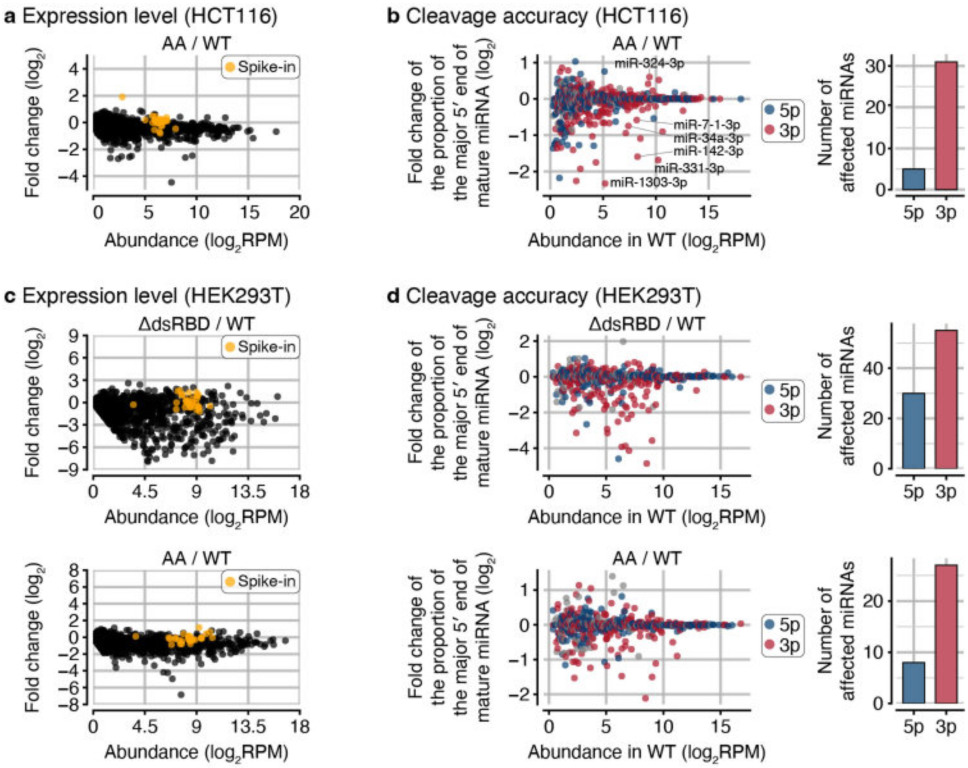


Extended Data Fig. 5 | See next page for caption.

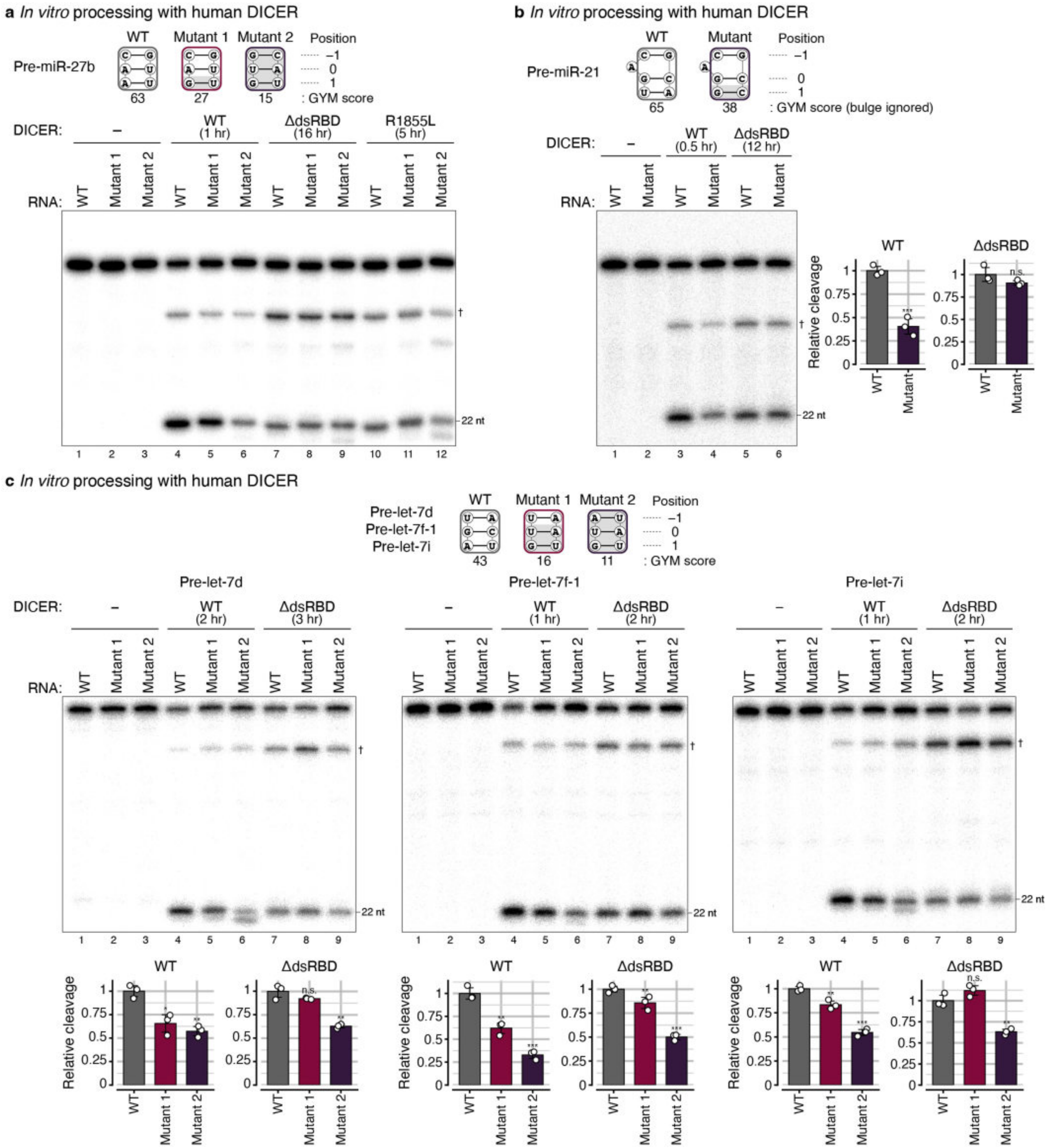
Extended Data Fig. 5 | The GYM motif affects efficiency and accuracy of DICER processing independently of TRBP. a–c, f, In vitro processing of pre-let-7a-1 variants by human DICER (a–b), human DICER and TRBP (c), or fly Dcr-1 (f). Substrates were radiolabeled at their 5′ ends. †, nicked products at the 3p positions. a, Lanes 6–10 are identical with those in Fig. 2a. b, Squares indicate mean (n = 2, independent experiments). c, Bars indicate mean ± SD (n = 3, independent experiments). ***$p < 0.001$ by two-sided Student's t test compared to GCm. d, DROSHA processing assay and miRNA abundance measurement of pre-miR-A1 variants in HEK293T cells. Left: Schematic outline of this experiment. Right top: Luciferase assay. Firefly luciferase signals were normalized to *Renilla* luciferase (Rluc) signals. Right bottom: miRNA levels measured by qRT-PCR. The TaqMan probe was designed to target the common sequence of variants. Bars indicate mean ± SD (n = 3, biological replicates). **$p < 0.01$, ***$p < 0.001$ by two-sided Student's t test compared to GCm. e,g, In vitro processing of duplex variants by human DICER. Cleavage products and their corresponding cleavage sites are marked with arrowheads. *, radiolabeled 5′ phosphate. g, The duplex had a base-pair at its terminus (marked in orange) so that the 5′ end cannot be incorporated into the 5′ pocket. For gel source data, see Supplementary Fig. 1.



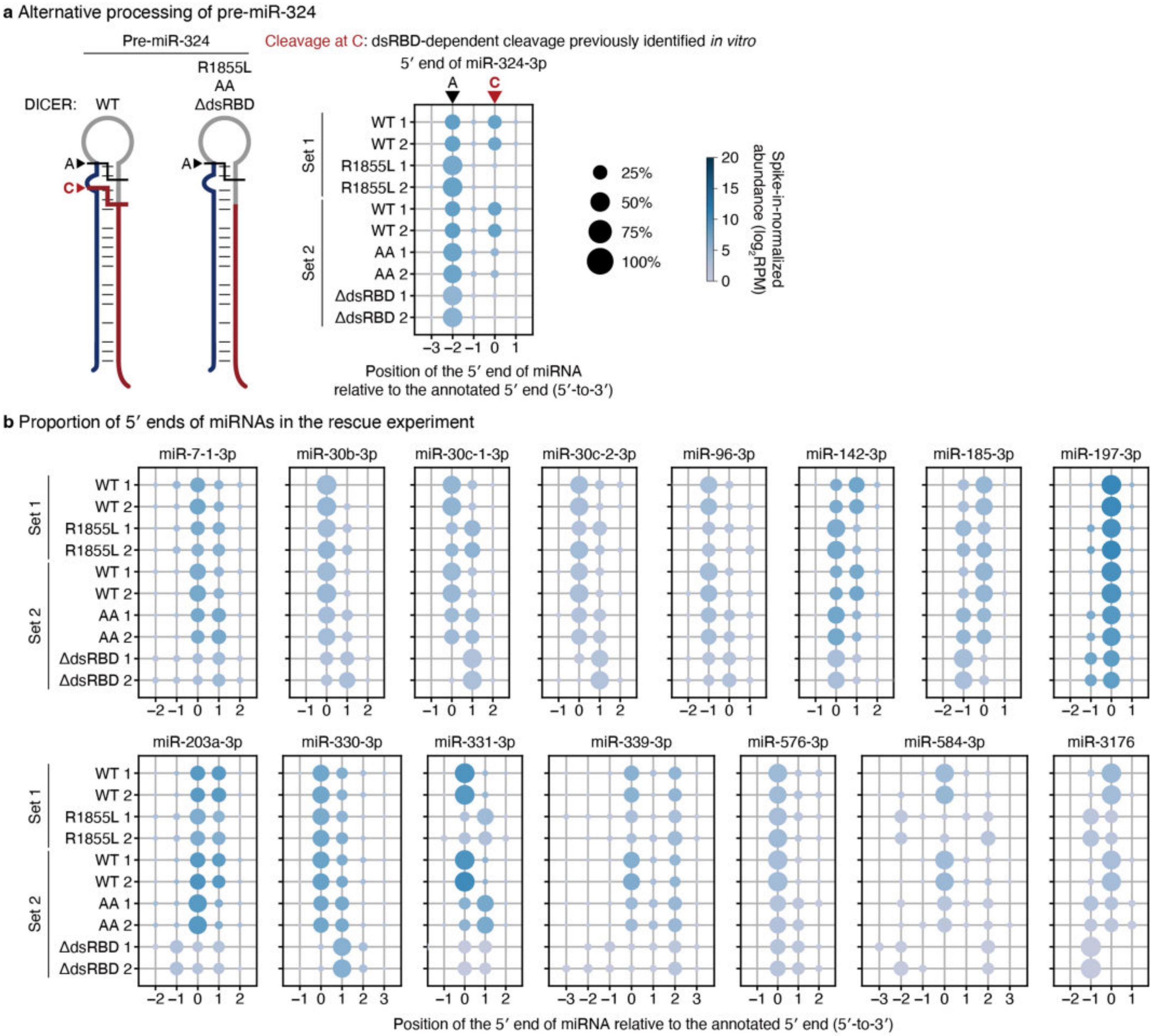
Extended Data Fig. 6 | R1855 and E1859 of the DICER dsRBD are important for recognition of the mismatch. a, In vitro processing of pre-let-7a-1 variants by human DICER ΔdsRBD with the indicated reaction time. b, Amino acid sequence alignment of dsRBDs of metazoan DICERs. c, In vitro processing of duplex variants by human DICER point mutants at the indicated position. Cleavage products and their corresponding cleavage sites are marked with arrowheads. *, radiolabeled 5′ phosphate. d,e, In vitro processing of pre-let-7a-1 variants by either DICER R1855L (d) or R1855A/E1859A (AA) (e) with the indicated reaction time. Bars indicate mean (n = 2, independent experiments) (d) or mean ± SD (n = 3, independent experiments) (e). *$p < 0.05$, ***$p < 0.001$ by two-sided Student's t test compared to GCm. †, nicked products at the 3p positions. f, In vitro processing of duplex variants by DICER AA mutant. The cleavage product and its corresponding cleavage site marked with the arrowhead are largely unaffected by the GYM motif variations, which contrasts the result from WT DICER shown in Fig. 2b, d. For gel source data, see Supplementary Fig. 1.



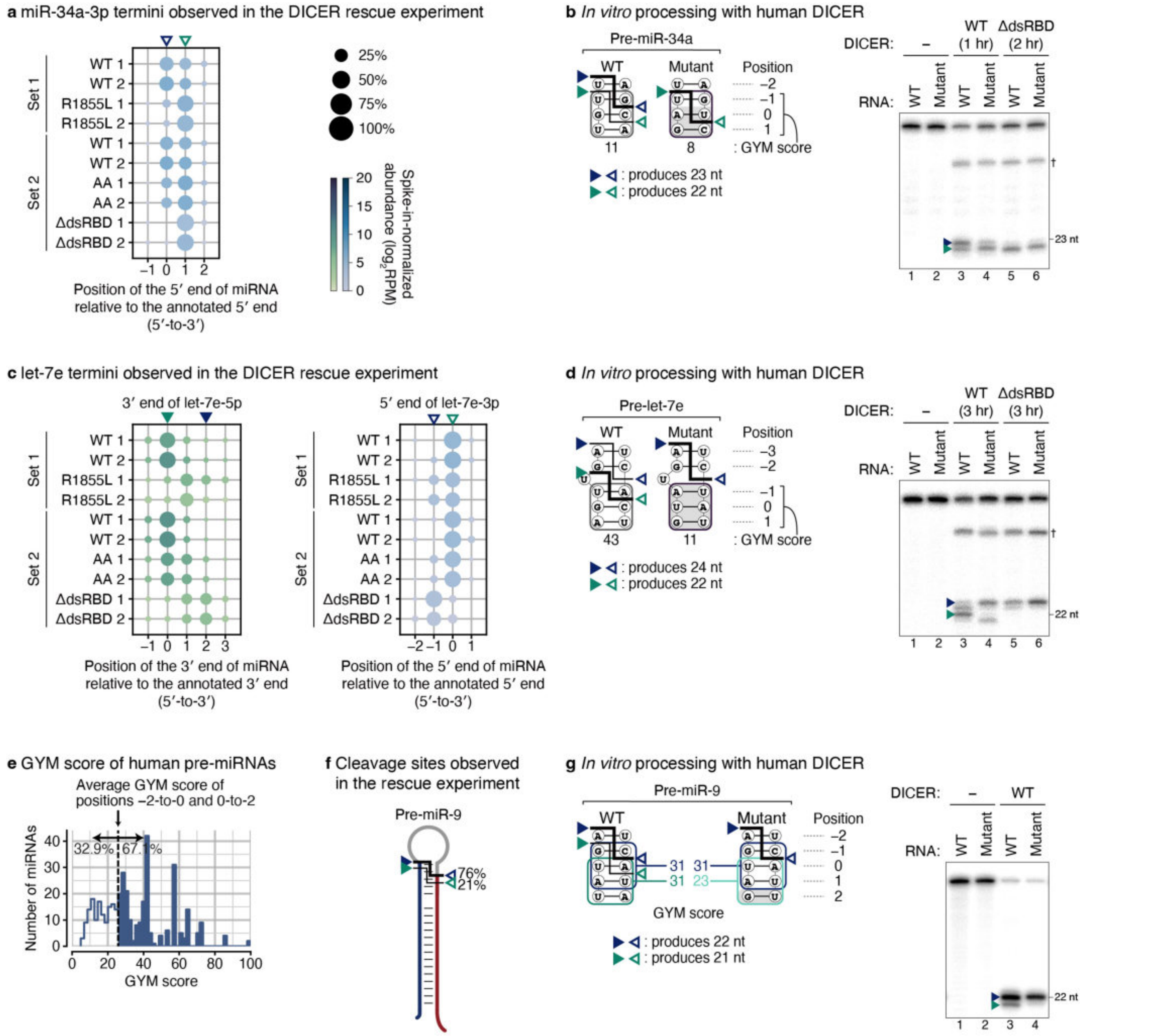
Extended Data Fig. 7 | Mutating the DICER dsRBD reduces efficiency and accuracy of DICER processing. a, c, Comparison of miRNA expressions in either HCT116 (a) or HEK293T (c). Spike-ins were used for normalization. RPM, reads per million. b, d, Comparison of cleavage accuracy in either HCT116 (b) or HEK293T (d). For a given miRNA, the most abundant 5′-isomiR was identified in the WT sample. Then the fold change of its proportions in each sample was measured as cleavage accuracy. Grey, unannotated strand. Bar graphs show the number of miRNAs whose major 5′-isomiR was significantly affected by the mutation ($p < 0.01$ by two-sided Student's t test).



Extended Data Fig. 8 | Processing of pre-miRNAs are regulated by the GYM motif recognized by the DICER dsRBD. a–c, In vitro processing of variants of pre-miR-27b (a), pre-miR-21 (b), and pre-let-7d/f-1/i (c) by either DICER WT or ΔdsRBD. Pre-miRNAs were radiolabeled at their 5′ end. Reactions were performed with different time points as indicated. For gel source data, see Supplementary Fig. 1. Bars indicate mean ± SD (n = 3, independent experiments). *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$ by two-sided Student's t test compared to the WT substrate. †, nicked products at the 3p positions.



Extended Data Fig. 9 | Examples of miRNAs whose DICER cleavage sites are affected by mutation of the DICER dsRBD. a–b, The usage of 5′ ends of miRNAs in the *DICER*-knockout HCT116 cells rescued with indicated DICER. The annotation in miRBase release 21 was used as a reference. Cleavage sites and their corresponding positions are marked with arrowheads. RPM, reads per million.



Extended Data Fig. 10 | The DICER dsRBD-GYM motif interaction plays a critical role in cleavage site decision of endogenous miRNAs. a, c, The usage of 5′ ends of miR-34a-3p (a) or 3′ ends of let-7e-5p and 5′ ends of let-7e-3p (c) in the *DICER*-knockout HCT116 cells rescued with indicated DICER. The annotations in miRBase release 21 were used as references. Corresponding positions of the major cleavage sites are marked with arrowheads. RPM, reads per million. b, d, In vitro processing of pre-miR-34a variants (b) or pre-let-7e variants (d) by either DICER WT or ΔdsRBD. Pre-miRNAs were radiolabeled at their 5′ end. Major cleavage products and their corresponding cleavage sites are marked with arrowheads. Reactions were performed with different time points as indicated because DICER ΔdsRBD has reduced activity. e, The GYM scores at the position −1 of human pre-miRNAs. miRNAs registered in miRGeneDB (n = 383) were included in this analysis. The dashed line indicates the average of GYM scores of the surrounding positions (−2 and 0). f, Alternative processing of pre-miR-9. Cleavage sites were inferred from 5′ ends of miR-9-3p in the *DICER*-knockout HCT116 cells rescued with DICER WT. Average proportions are indicated at the corresponding cleavage sites marked with arrowheads. g, In vitro processing of pre-miR-9-1 by DICER. The GYM score for each window (grey and colored boxes) is shown. Pre-miRNAs were radiolabeled at their 5′ end. Major cleavage products and their corresponding cleavage sites are marked with arrowheads. For gel source data, see Supplementary Fig. 1.

nature portfolio

Corresponding author(s): V. Narry Kim

Last updated by author(s): Dec 21, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Sequencing data were collected by the software within the NovaSeq 6000.
Data analysis	Preprocessing for the sequencing data was done with cutadapt 1.8.3, FASTX Toolkit 0.0.13.2, bwa 0.7.10-r789, bedtools v2.25.0. Custom analysis codes are available at https://github.com/haedongkim615/dicer_gym_motif. Visualization of structural models were performed using and PyMol (version 2.4.1). Pre-miRNA secondary structures were obtained using Fold in RNAstructure version 6.3 with default settings.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The massively parallel assay data and rescue data were deposited to the GEO repository (accession number: GSE202535 and GSE215866). Other structural models

cited in this study for analysis (5ZAL and 2EZ6) are also accessible on PDB. The Cancer Genome Atlas (TCGA) data for DICER1 gene was accessed at cBioPortal (https://www.cbioportal.org).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Not applicable
Population characteristics	Not applicable
Recruitment	Not applicable
Ethics oversight	Not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Described in the corresponding figure legends. No sample-size calculation was performed. For in vitro processing assays, we have performed independent experiments, which is sufficient to estimate the variation.
Data exclusions	No data was excluded.
Replication	Experiments were done with replicates as indicated in the figure legends. Independent methods, such as massively parallel assays and in vitro DICER processing, were in agreement.
Randomization	This project does not involve allocation into experimental groups.
Blinding	Blinding was not relevant to this study. The results from in vitro processing assays are visualized by urea-PAGE as individual bands that are quantified using software (i.e., ImageJ), requiring no subject interpretation or judgement.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	☐	Antibodies
☐	☒	Eukaryotic cell lines
☒	☐	Palaeontology and archaeology
☒	☐	Animals and other organisms
☒	☐	Clinical data
☒	☐	Dual use research of concern

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	HCT116 cell lines (WT and DICER KO) were obtained from ToolGen (Kim et al., PNAS, 2016). HEK293T cell lines (WT and DICER KO) were gifted from Bryan Cullen's lab (Bogerd et al., RNA, 2014).
Authentication	Authenticated using the ATCC short tandem repeat profiling. Last performed on Mar 25th, 2022.
Mycoplasma contamination	Cell lines were confirmed mycoplasma negative prior to use.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used in this study.

Structure of the human DICER–pre-miRNA complex in a dicing state

https://doi.org/10.1038/s41586-023-05723-3

Received: 30 May 2022

Accepted: 14 December 2022

Published online: 22 February 2023

Check for updates

Young-Yoon Lee[1,2,4], Hansol Lee[2,3,4], Haedong Kim[1,2,4], V. Narry Kim[1,2]✉ & Soung-Hun Roh[2,3]✉

Dicer has a key role in small RNA biogenesis, processing double-stranded RNAs (dsRNAs)[1,2]. Human DICER (hDICER, also known as DICER1) is specialized for cleaving small hairpin structures such as precursor microRNAs (pre-miRNAs) and has limited activity towards long dsRNAs—unlike its homologues in lower eukaryotes and plants, which cleave long dsRNAs. Although the mechanism by which long dsRNAs are cleaved has been well documented, our understanding of pre-miRNA processing is incomplete because structures of hDICER in a catalytic state are lacking. Here we report the cryo-electron microscopy structure of hDICER bound to pre-miRNA in a dicing state and uncover the structural basis of pre-miRNA processing. hDICER undergoes large conformational changes to attain the active state. The helicase domain becomes flexible, which allows the binding of pre-miRNA to the catalytic valley. The double-stranded RNA-binding domain relocates and anchors pre-miRNA in a specific position through both sequence-independent and sequence-specific recognition of the newly identified 'GYM motif'[3]. The DICER-specific PAZ helix is also reoriented to accommodate the RNA. Furthermore, our structure identifies a configuration of the 5′ end of pre-miRNA inserted into a basic pocket. In this pocket, a group of arginine residues recognize the 5′ terminal base (disfavouring guanine) and terminal monophosphate; this explains the specificity of hDICER and how it determines the cleavage site. We identify cancer-associated mutations in the 5′ pocket residues that impair miRNA biogenesis. Our study reveals how hDICER recognizes pre-miRNAs with stringent specificity and enables a mechanistic understanding of hDICER-related diseases.

Small regulatory RNAs serve as guide molecules in RNA interference (RNAi) by inducing translational repression and destabilization of the cognate mRNAs[4–6]. Central to the pathway is the ribonuclease (RNase) III enzyme Dicer, which cleaves long dsRNAs or short hairpin RNAs to generate small RNAs of 21–25 nucleotides (nt) in length[1,2]. Dicer homologues are found throughout eukaryotes and show substantial diversity in their substrate specificity and mechanism of action. Some Dicer proteins are specific to long dsRNAs, as observed in structural and biochemical studies on *Giardia* Dicer, fly Dicer-2 (Dcr-2) and plant Dicer-like proteins[7–12]. By contrast, other homologues, such as fly Dicer-1 (Dcr-1), are highly selective to hairpin-shaped pre-miRNAs[13]. hDICER can cleave both types of substrate, with a clear preference for short hairpins over long dsRNAs[14,15].

hDICER recognizes several features of its substrates: a dsRNA stem of approximately 22 bp; a 2-nt 3′ overhang; and a flexible loop next to the cleavage site[7,12,13,16–19]. The flexible loop is known to be sensed by the helicase domain[13,14], whereas the 5′ phosphorylated end and the 3′ overhang are recognized by basic pockets in the platform and the PAZ (Piwi–Argonaute–Zwille) domains, respectively[16,17,20]. By anchoring the termini, hDICER can act as a 'molecular ruler' to measure around 22 nt away from the 5′ end ('5′ counting rule') and 3′ end ('3′ counting rule') of the substrate[7,12,16,17]. In addition, the GYM motif at the cleavage site enables the cleavage site to be precisely determined (see the partner paper to this one[3]). However, the structural basis of the substrate specificity of hDICER remains largely unknown, owing to the lack of an active-state structure.

Early electron microscopy (EM) analyses of hDICER revealed its overall L-shape[21–23]. Crystal structures of a partial fragment containing the platform–PAZ domain showed the 5′ and 3′ pockets, which recognize the respective ends of a small RNA duplex[20]. A more recent cryo-electron microscopy (cryo-EM) study showed the overall topology of full-length hDICER in the apo and RNA-bound states[24]. However, in this structure, the pre-miRNA is situated distant from the catalytic valley, probably representing a 'pre-dicing' state. Thus, we still lack a structural understanding of how hDICER recognizes pre-miRNAs in an active state.

Here we aimed to determine the cryo-EM structure of hDICER with pre-miRNA in a cleavage-competent state. The structure reveals dynamic spatial rearrangements of several domains of hDICER during the transition to a catalytic state, and explains how hDICER selects its substrates with specificity.

[1]Center for RNA Research, Institute for Basic Science (IBS), Seoul, Republic of Korea. [2]School of Biological Sciences, Seoul National University, Seoul, Republic of Korea. [3]Institute of Molecular Biology and Genetics, Seoul National University, Seoul, Republic of Korea. [4]These authors contributed equally: Young-Yoon Lee, Hansol Lee, Haedong Kim. ✉e-mail: narrykim@snu.ac.kr; shroh@snu.ac.kr

Structural determination

To reconstitute the enzyme–substrate complex, we used mono-uridylated pre-let-7a-1, which has an optimal 2-nt 3′ overhang, stem length and terminal loop. Meanwhile, in our concurrent study[3], we performed massively parallel assays using more than one million pre-miRNAs with random sequences near the cleavage sites. An analysis of the top variants revealed that nucleotide sequences at positions −1, 0 and 1 relative to the 3p DICER cleavage site were enriched with a paired guanine (G), a paired pyrimidine (Y) and a mismatched cytosine or adenine (M), respectively. We therefore termed this motif the GYM motif. The GYM motif robustly enhances dsRNA processing at a specific site, suggesting that it has a role in the catalytic step. The mismatched 'M' is particularly important for the decision of cleavage site. To design an optimal substrate for structural determination, as well as to obtain mechanistic insights into GYM-motif-mediated processing, we incorporated the highest-scoring GYM motif (5′-CGC/GCC-3′) into pre-let-7a-1 (referred to as pre-let-7a-1^{GYM}) (Extended Data Fig. 1a), which substantially increased the processing rate compared to the wild-type sequence.

We overexpressed and purified the full-length hDICER protein, which cleaved the pre-let-7a-1^{GYM} precisely, yielding the expected 22-nt fragment (Extended Data Fig. 1b–d). Previous cryo-EM structures of hDICER were determined with its accessory protein, TRBP, which exhibits flexible and heterogenous conformations[24]. As the vast majority of human miRNAs, including let-7, do not require TRBP both in vitro and in cells[16,25,26], we set out to determine the cryo-EM structure of hDICER alone so as to reduce the structural heterogeneity.

We first validated the architecture of the purified apo-hDICER at 4.0 Å resolution by cryo-EM (Extended Data Fig. 2a–f and Extended Data Table 1). This structure exhibits a compact hatchet-like (or L-shaped) architecture[21–23] with mostly globular domains, as observed in the previous cryo-EM structure of the hDICER–TRBP complex[24]. Our map clearly shows overall domains with defined helical features (Extended Data Fig. 2g,h). The helicase domain shows lower local resolution, suggesting its intrinsically flexible behaviour (Extended Data Fig. 2f). Next, to build a model of apo-hDICER, we introduced the cryo-EM model of hDICER from hDICER–TRBP into our map and refined it by flexible fitting[24]. Our model covers the three-dimensional (3D) position of most folded domains and partial interdomain loops, including DUF283 linker (D-linker), which connects DUF283 to the platform domain (Extended Data Fig. 2g,h). Notably, D-linker, which was mostly undetermined in the previous hDICER–TRBP structure, forms a short β-sheet with the N-terminal region of RNase III domain a (RIIIDa).

Next, we reconstituted the hDICER–pre-miRNA complex (Extended Data Fig. 1e–g) and were able to reconstruct the map of the complex at 3.0-Å resolution using a cryo-EM dataset of around 1,210,800 particle images (Extended Data Fig. 3a–f and Extended Data Table 1). In this map, hDICER embraces the helix of pre-miRNA within the catalytic centre of the enzyme through an extensive surface contact (Fig. 1a–c). Our map shows the side-chain details of the platform, PAZ, RIIIDa and RIIIDb domains and the secondary structural features for the double-stranded RNA-binding domain (dsRBD) and interdomain linkers. We could not identify the featured densities from helicase (residues 1–564), DUF283 (residues 590–714) and several loops in RIIID domains (residues 1392–1545 and 1595–1684) (Fig. 1a,b). We referenced the model of apo-hDICER and pre-let-7a-1 structures and built an atomic model of hDICER–pre-let-7a-1^{GYM} (Extended Data Fig. 3g–i).

Overall structure in a dicing state

The structure of the hDICER–pre-let-7a-1^{GYM} complex shows that the pre-miRNA is fully docked and poised for cleavage within the catalytic centre that is formed by intramolecular dimerization of RIIIDa and RIIIDb (Fig. 1d). The catalytic centre has two clusters of conserved acidic amino acid residues in the RIIIDa (E1316, D1320, D1561 and E1564) and RIIIDb (E1705, D1709, D1810 and E1813) (Fig. 1e,f and Extended Data Fig. 4a–c). Our map also shows extra densities in the catalytic core, which are attributed to calcium ions used to substitute magnesium ions to prevent hydrolysis[27]. The calcium ions are situated near the oxygen atoms of the scissile phosphodiester bonds in the 5p strand (between 22U and 23U) and the 3p strand (between 51G and 52C) (Fig. 1e,f), the position of which coincides with the actual cleavage sites of pre-let-7a-1. This spatial arrangement is highly homologous to that of other RNase III type enzymes, including human DROSHA and *Aquifex aeolicus* RNase III (Aa RNase III)[28–30] (Extended Data Fig. 4d,e).

We could build the 3D model for most of the pre-miRNA, including the stem region and the additional 4 nt and 6 nt beyond the cleavage sites in the 5p and 3p strands, respectively (Fig. 1c,d). The rest of the terminal loop could not be modelled, affirming the flexible nature of the terminal loop. Our structure shows considerable contacts between hDICER and pre-miRNA, with a total buried surface area of 10,290.8 Å^2 (Extended Data Fig. 4f). The root-mean-square deviation (RMSD) between the apo and dicing states was 2.9 Å, mostly accounting for differences in the dsRBD and PAZ domain, with RMSDs of 17.7 Å and 13.6 Å, respectively (Extended Data Fig. 4g). Compared to the pre-dicing state[24], the dicing-state structure shows large differences both in protein domain organization and RNA interaction. Note that in the pre-dicing state, there is only limited interaction with pre-miRNA, mainly through its termini and loop (with a total buried surface area of 3,206.9 Å^2) (Extended Data Fig. 4f).

One of the prominent changes observed during the transition between apo and dicing states was the fade-out of the density of helicase domain. In the apo state structure, there appear to be interdomain interactions among the helicase domain, dsRBD and RIIIDb (Extended Data Fig. 5a). These interactions are likely to support the overall architecture of these domains in a stable 'closed' conformation. However, superposition of the apo-hDICER structure, into which the helicase domain could be modelled, shows a steric clash between the helicase domain and the pre-miRNA loop (Extended Data Fig. 5b). Consistently, the N-terminal helicase and DUF283 domains exhibit substantial flexibility in a dicing state in our analysis (Fig. 1b). We observed the same result with pre-miR-3121^{GYM} that has a small loop (11 nt), suggesting that the helicase domain becomes flexible generally during the dicing step regardless of the terminal loop size (Extended Data Fig. 5c). Further in-depth particle classification of around 4,000 particles revealed the extended map of the helicase domain, dislocated from other domains (Extended Data Fig. 5d). To test whether this missing density is due to chemical integrity, we performed another cryo-EM imaging of the same specimens, after incubating with 2 mM $MgCl_2$, which allowed the cleavage reaction (Extended Data Fig. 5e). We obtained particles at multiple structural states including the dicing state (25%), some intermediates (51%) and apo-like structures (34%) (Extended Data Fig. 5f). This structural analysis indicates that the helicase domain is chemically intact and that the structural heterogeneity of the helicase domain is induced transiently in the dicing state. After the cleavage reaction, the protein returns to the original apo state conformation. Collectively, these analyses suggest that the conformational change in the helicase domain is required to allow productive interaction with the substrate. These observations contrast with structures of fly Dcr-2—specialized in the short interfering RNA (siRNA) pathway—that showed a fixed orientation of the helicase domain, which is engaged in ATP-dependent translocation of the long dsRNA[10,11].

Stem recognition by the dsRBD and RIIID

Near the catalytic sites in the upper stem of pre-miRNA, we observed a pronounced movement of the C-terminal dsRBD (Fig. 2a), mediated by the flexible linker that connects to the RIIIDb. This RNA-induced conformational switching orients the dsRBD away from the inner core

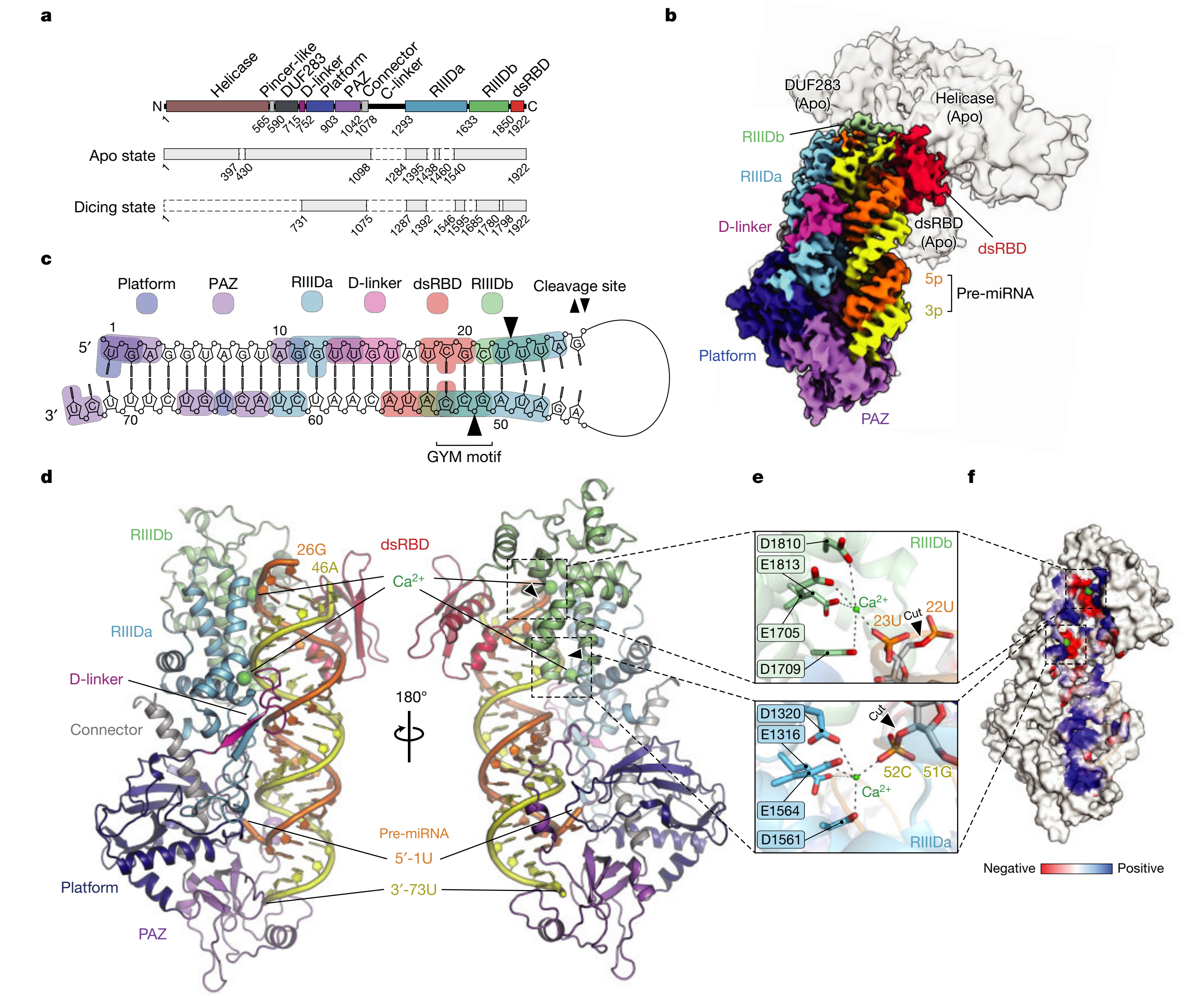


Fig. 1 | Cryo-EM structure of hDICER in complex with a pre-miRNA in a dicing state. a, Domain organization of hDICER. Schematics for the apo and dicing states indicate amino acid residues that are included (solid lines) or not included (dashed lines) in the model. C-linker, connector linker; DUF, domain of unknown function. **b**, Cryo-EM map of hDICER in a dicing state (colour, 3.0 Å) overlaying that of hDICER in an apo state (grey, 4.0 Å). **c**, Protein–RNA interactions in the dicing state at the domain level. Sequences of pre-let-7a-1^{GYM} that are not included in the model are not shown. **d**, Overall structure of the hDICER with a pre-miRNA in a cleavage-competent state. Black arrowheads point to the DICER cleavage sites. **e**, Magnified views of the catalytic sites in the RIIIDa and RIIIDb domains. **f**, Electrostatic potential surface model of the catalytic valley along the protein–RNA interface.

of the catalytic valley and relieves the steric clash between the dsRBD and dsRNA (Extended Data Fig. 6a) so as to permit dsRNA recognition. With respect to its original position in the apo and pre-dicing states, the dsRBD swings about 12.6 Å and 16.5 Å, respectively (Fig. 2a and Extended Data Fig. 6a).

Of note, close to the dsRBD–dsRNA interface, we observed a large change in the helical structure of dsRNA, which deviates from the ideal A-form dsRNA structure (Fig. 2b). This conformational distortion expands the width of the major groove of pre-let-7a-1^{GYM} to 15.6 Å, compared with 8.0 Å in the ideal A-form RNA. The successive major and minor grooves across the region are sandwiched between dsRBD and RIIIDa, and form extensive interactions with both domains (Fig. 2c), suggesting a possible basis for the local distortion in the dsRNA structure. Similar observations were made in the high-resolution cryo-EM structure of *Arabidopsis* DCL-3 (AtDCL3), in complex with a pre-siRNA[9] (Extended Data Fig. 6b), implying that the conformational distortion in the dsRNA helix is not specific to the pre-let-7a-1 sequence, but induced by protein–RNA interactions that are unique to a certain group of Dicer homologues.

In addition to the dsRBD, the RIIIDa and RIIIDb domains wrap around the RNA, forming extensive electrostatic interactions with the upper stem region of the pre-miRNA. As well as the contacts at the catalytic core of the DICER cleavage sites, the RIIIDa, situated on the opposite face to the dsRBD-binding site, makes tight interactions with the RNA in the minor groove (Fig. 2c). We observed that α-helices 2 and 3 of RIIIDa potentially interact with the ribose sugars and internucleotide phosphate groups (Fig. 2d). The symmetrically located α-helices 2 and 3 in the RIIIDb may also participate in dsRNA recognition (Fig. 2e).

Interacting with the distorted dsRNA, the dsRBD adopts a canonical αβββα topology to cover the dsRNA across the minor and major grooves (Fig. 2c). Basically, the reoriented dsRBD interacts with the RNA backbone through its mostly basic patch on the surface (Extended Data Fig. 6c). For instance, in the major groove, the positively charged residues contact the RNA backbone near 19C (Fig. 2f and Extended Data

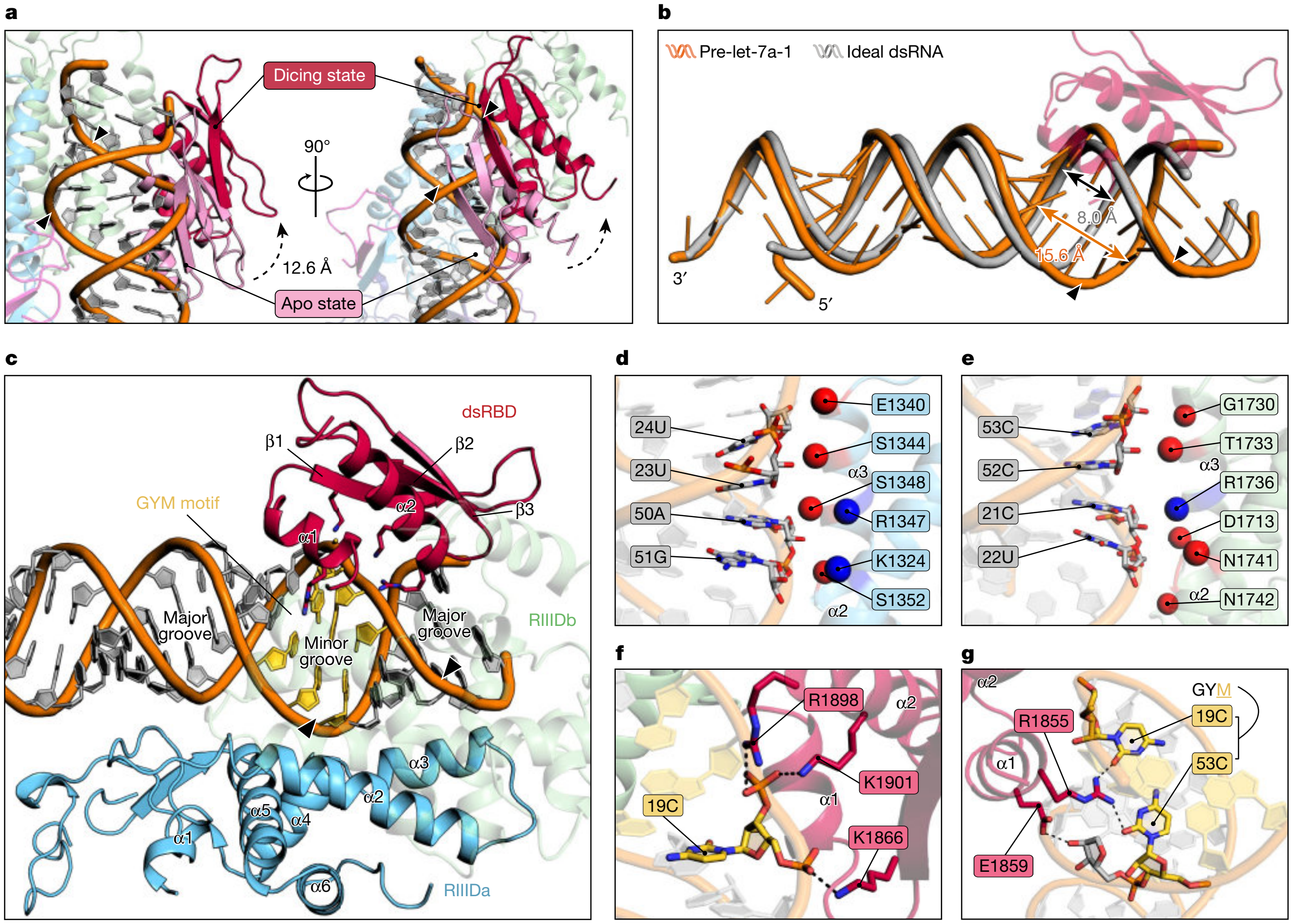


Fig. 2 | Sequence-specific and non-specific binding to RNA by the dsRBD and RIIID domains. a, Conformational change in the dsRBD during the transition from an apo state to a dicing state. Black arrowheads near RNA backbones indicate cleavage sites. **b**, Comparison between the structures of ideal A-form dsRNA helix and pre-let-7a-1^{GYM}. **c**, Protein–RNA interactions near the cleavage sites in the minor groove. **d**, Protein–RNA interactions in the interface between RIIIDa and the RNA backbone. Residues that may engage in electrostatic interactions with the RNA backbone are displayed as balls. **e**, Protein–RNA interactions in the interface between RIIIDb and the RNA backbone. Residues that may engage in electrostatic interactions with the RNA backbone are displayed as balls. **f**, Non-sequence-specific interactions between the dsRBD and the RNA phosphate backbone. **g**, Sequence-specific interactions between the dsRBD and the C–C mismatch of the GYM motif.

Fig. 6d). In the minor groove, however, we observed a sequence-specific interaction between the dsRBD and the RNA. The α-helix 1 of the dsRBD is situated in the vicinity of the GYM motif (Fig. 2c,g), the *cis*-acting element that markedly improves the fidelity of processing (see the partner paper[3]). An arginine residue (R1855) in α-helix 1 protrudes into the minor groove and may form hydrogen bonds with the C–C mismatch (Fig. 2g and Supplementary Video 1). This is consistent with the observation that mutating this arginine residue abolishes the effect exerted by the mismatch[3]. Thus, in contrast to previous papers that suggest an auxiliary function of the hDICER dsRBD[31,32], our structure provides the structural basis for its predominant role in the selection of the cleavage site, which can even override the effects of 5′ and 3′ counting mechanisms[3]. Together, our data indicate that the dsRBD and RIIID domains anchor the upper region of the pre-miRNA and induce local distortion of the RNA, which facilitates the recognition of not only the RNA backbone but also the GYM motif.

Role of the PAZ helix in a dicing state

A previous structural study on a hDICER fragment containing the platform–PAZ domain showed a knob-like protrusion with a small α-helical segment within the PAZ domain, which is specifically found in Dicer[20]. This 'PAZ helix' (also known as hDICER-specific helix) separates the 5′ and 3′ pockets and orients the bound RNA away from the surface of hDICER, which is thought to occur in a product-release state and/or pre-dicing state[20] (Extended Data Fig. 7a, middle). This helix, however, may be dynamic given that an additional 'melted' conformation of the PAZ helix was observed in the platform–PAZ–small RNA duplex complex[20] (Extended Data Fig. 7a, right). Indeed, we found a large conformational change in the PAZ helix resulting in a tilt angle of around 54° from its position in the pre-dicing state (Fig. 3a). The PAZ helix is consequently located near the lower stem region of pre-miRNA (Fig. 3a and Supplementary Video 2). Together, our findings show that the spatial rearrangement of the PAZ helix is necessary to allow the pre-miRNA to be aligned parallel to the catalytic valley for subsequent cleavage (Extended Data Fig. 7b).

In addition, this conformational change puts the short stretch of positively charged amino acids ($_{1019}$KRKAK$_{1023}$) in the vicinity of the negatively charged backbone of the 3p strand (Fig. 3b). To assess the importance of the observed interaction between the PAZ helix and the dsRNA, we introduced mutations by replacing the positively charged residues with five glutamate (E5) or alanine (A5) residues, or by deleting the helix (ΔPAZh) (Fig. 3c). The PAZ-helix-mutant proteins were purified (Extended Data Fig. 1b,c) and incubated with pre-let-7a-1 to

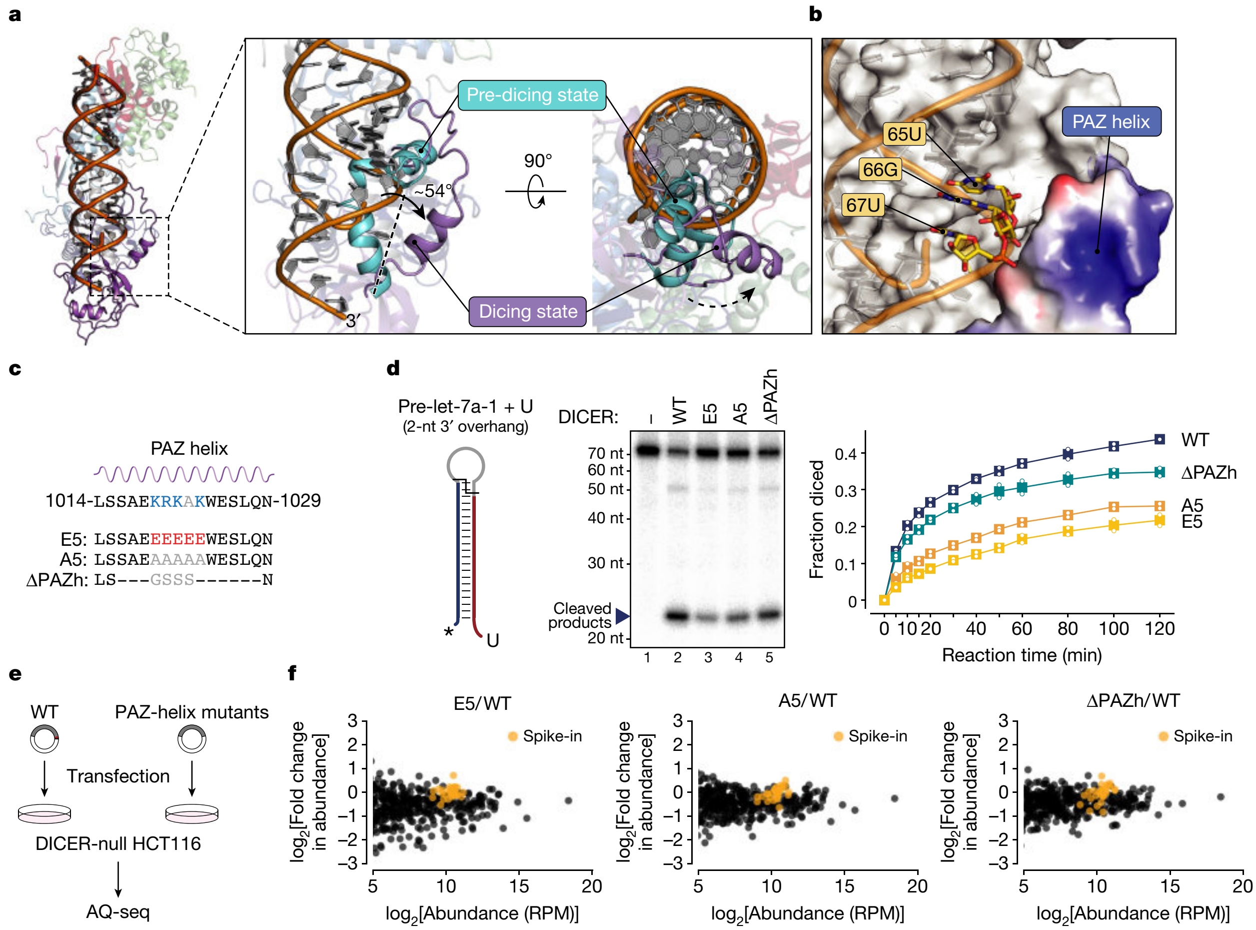


Fig. 3 | The PAZ helix reorganizes to accommodate pre-miRNA in a dicing state. a, Conformational change of the PAZ helix between a dicing state (this study) and a pre-dicing state[24]. **b**, Electrostatic interactions between the positively charged PAZ helix and the negatively charged RNA phosphate backbone. **c**, Sequences of DICER mutants. **d**, In vitro processing of mono-uridylated pre-let-7a-1. Relative cleavage was calculated by quantifying the band intensity (1 − uncleaved/input). Squares indicate mean ($n = 2$, independent experiments). Asterisk indicates radiolabelled 5′ phosphate. For gel source data, see Supplementary Fig. 1. **e**, Schematic outline of the rescue experiment ($n = 2$, biological replicates). **f**, Comparison of miRNA abundance. Spike-ins were used for normalization. RPM, reads per million.

quantitatively measure their effects on the cleavage efficiency in vitro. The mutant proteins showed reduced cleavage efficiencies, regardless of the length of the 3′ overhang (Fig. 3d and Extended Data Fig. 7c,d). PAZ helix^{E5} showed a more severe effect than PAZ helix^{A5}, presumably owing to the electrostatic repulsive forces created between the PAZ helix and the RNA backbone. Notably, the deletion of the PAZ helix led to a modest but consistent reduction in cleavage efficiency. This result, together with the structural observations, implies that the PAZ helix has an autoinhibitory effect on the transition to a dicing state besides its contribution to RNA-binding affinity, once the dicing state is achieved.

We next sought to investigate the role of the PAZ helix in miRNA biogenesis by transiently expressing the mutant DICER in *DICER*-knockout HCT116 cells, and then performed Accurate Quantification by Sequencing (AQ-seq), which reliably profiles cellular miRNAs[33] (Fig. 3e). The mutations resulted in a global reduction in the abundance of miRNAs (Fig. 3f), corroborating the in vitro results. Our data collectively suggest that the conformational change in the PAZ helix and its subsequent interaction with the RNA backbone are important for pre-miRNA processing.

Architecture of the 5′ and 3′ pockets

Consistent with the idea that hDICER recognizes the pre-miRNA termini for accurate processing, our structure shows both 5′ and 3′ ends stably anchored within the platform and PAZ domains, respectively (Fig. 4a–c and Extended Data Fig. 8a,b). In the 3′ pocket, the last phosphodiester linkage makes close interactions with a cluster of four conserved tyrosine residues (Y936, Y971, Y972 and Y976) and an arginine residue (R937) through potential hydrogen bonding, which is in line with previous reports[20,24] (Fig. 4c).

The 5′ end of pre-miRNA is in a unique kinked conformation (Fig. 4a), which is very different from the previous structures of hDICER with a small RNA duplex or a pre-miRNA in the pre-dicing state and other Dicer homologues in the dicing state[9,20,24] (Extended Data Fig. 8c–g). The conformation in our structure allows the 5′ monophosphate to be inserted into the 5′ pocket and possibly interact through hydrogen bonds with the main-chain amide and the amine group of two arginine residues, R996 and R1003, respectively (Fig. 4b and Extended Data Fig. 8a). In addition, the 5′ base unexpectedly flips out to interact with a cluster of three arginine residues—R788, R790 and R821—that make hydrogen bonds with the base (Fig. 4b). These results collectively suggest that the 3′ pocket is conserved whereas the 5′ pocket may have emerged more recently to meet the needs of individual Dicer homologues with different substrate types.

Cancer-associated 5′ pocket mutations

The 5′ counting mechanism is important for miRNA biogenesis because it ensures accurate and efficient processing regardless of the frequent

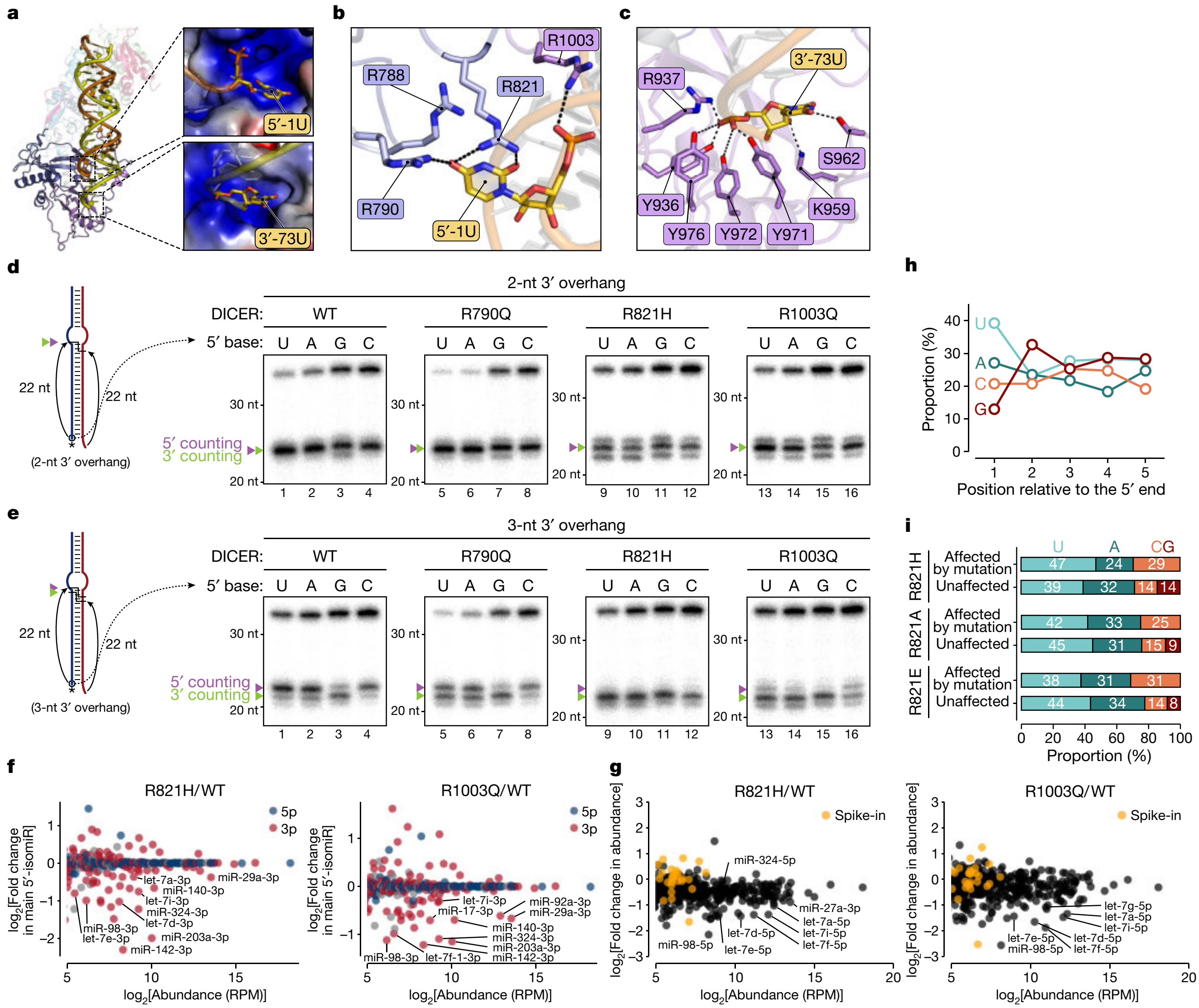


Fig. 4 | Cancer-associated mutations of the 5′ pocket affect miRNA biogenesis. a, The 5′ and 3′ ends of pre-miRNA anchored in the basic 5′ and 3′ pockets in the platform and PAZ domains, respectively. **b**, The 5′-end recognition by the 5′ pocket through hydrogen bonding. **c**, The 3′-end recognition by the 3′ pocket through hydrogen bonding. **d**,**e**, Cancer-associated mutations in the 5′ pocket impair 5′ counting. In vitro processing of duplex RNAs with either a 2-nt (**d**) or a 3-nt (**e**) 3′ overhang and a varying sequence at the 5′ end. The nucleotide in the 3p strand opposite to the varying sequence is A. Asterisk indicates radiolabelled 5′ phosphate. For gel source data, see Supplementary Fig. 1. **f**, Changes in cleavage accuracy, estimated with the fold change of the proportion of the major 5′-isomiR. For a given miRNA, the most abundant 5′-isomiR was identified in the wild-type (WT) sample. Grey, unannotated strand. **g**, Change in abundance of endogenous miRNAs. Spike-ins were used for normalization. **h**, Nucleotide composition of human pre-miRNAs. The proportion of each base at the indicated position relative to the 5′ end of pre-miRNAs is shown. miRNAs whose 5p is registered in MirGeneDB (https://mirgenedb.org/) ($n = 502$) were included in this analysis. **i**, The 5′ terminal base identity of pre-miRNAs. Compare the miRNA groups whose major 5′-isomiRs are affected by the R821 mutation with those that are unaffected. Affected miRNAs ($P < 0.05$, $\log_2$-transformed fold change > 0.6 or < −0.6); unaffected miRNAs ($P > 0.05$); P value by two-sided Student's t-test compared to wild type.

3′ end trimming and tailing in cells[16]. Of note, three amino acids in the 5′ pocket are mutated in diverse cancer types, according to cancer databases including The Cancer Genome Atlas: R790Q (in colon adenocarcinoma); R821H (in colorectal adenocarcinoma); and R1003Q (in uterine endometrioid carcinoma and rectal adenocarcinoma)[34,35].

In vitro, the R821H and R1003Q mutants show defects in cleavage-site selection; they produced multiple products (21–23 nt) from dsRNA with a canonical 2-nt 3′ overhang (Fig. 4d, lanes 9–16). Note that wild-type hDICER cleaves this substrate homogeneously because the 5′ and 3′ counting mechanisms corroborate to ensure the generation of a 22-nt product (Fig. 4d, lanes 1–4). To differentiate the 5′ counting and 3′ counting mechanisms more clearly, we next used dsRNA with a 3-nt 3′ overhang (Fig. 4e). In this assay, all three mutants showed a considerable reduction in the 5′ counting capability, validating our structural observation that these residues are in close contact with the 5′ end of RNA. The R821H mutation as well as two other mutations of R821 (R821E and R821A) resulted in a complete loss of 5′ counting (Fig. 4e and Extended Data Fig. 8h).

To interrogate the effects on miRNA biogenesis, we performed rescue experiments with wild-type and mutant hDICER (Extended Data Fig. 8i). We observed marked alterations in the DICER cleavage sites of many miRNAs (Fig. 4f and Extended Data Fig. 8j), as indicated by the 5′ end of 3p miRNAs, which is determined at the DICER processing step (Extended Data Fig. 8k). The DROSHA processing sites (5′ ends of 5p mature miRNAs) were largely unaffected, as expected. The most notable examples are let-7-3p and miR-324-3p, which are dependent on the 5′

pocket[16,36] (Fig. 4f and Extended Data Fig. 8j). Other 3p miRNAs (such as miR-29a, miR-92a, miR-140, miR-142 and miR-203a) also showed substantial changes in the processing sites, which alter the miRNA isoform landscape and their target repertoire.

Moreover, the abundance of many miRNAs was decreased by the mutations (Fig. 4g). Notably, the mutations led to a significant reduction in the levels of tumour-suppressive let-7 family members—let-7a, let-7d, let-7e, let-7f, let-7i and miR-98—which were shown to be downregulated in various cancers[37,38]. Our results indicate the key role of the 5′-end recognition and suggest the oncogenic potential of the 5′ pocket mutations.

Role of the 5′ base

In the 5′ pocket, R821 potentially forms hydrogen bonds with the O2 carbonyl of the 5′ terminal uracil (Fig. 4b), hinting at a previously unknown base specificity. To examine the base specificity, we modelled other bases—cytosine, adenine and guanine—into the corresponding position within the 5′ pocket structure (Extended Data Fig. 9a). The guanine, unlike other bases, has a 2-amino group that sterically overlaps with the guanidino group of R821, prompting us to experimentally test the effect of the 5′ terminal sequence on the 5′ counting rule.

In vitro, dsRNAs with 5′-U, 5′-C and 5′-A bases are cleaved predominantly by the 5′ counting mechanism, whereas the substrate with 5′-G is cleaved mainly by the 3′ counting mechanism (Fig. 4e, lanes 1–4). Even in the presence of a 2-nt 3′ overhang, the dsRNA substrate with 5′-G was imprecisely cleaved, yielding 21–22-nt products (Fig. 4d, lanes 1–4), which is consistent with a previous observation that defects in 5′ counting give rise to multiple products[16]. These observations are not explained by the previous notion that pre-miRNAs with high thermodynamic stability at the 5′ terminus do not follow the 5′ counting rule[16], because among our substrates, only the dsRNA starting with 5′-U forms a base pair at the terminus. Together, these data suggest that 5′-G may be inefficiently inserted to the 5′ pocket, compromising the 5′-end recognition.

To investigate whether R821 is involved in the recognition of the 5′ end, we performed the same in vitro processing assays with R821 mutants (Extended Data Fig. 1b,c). The R821 mutants completely lost the 5′ counting capability, and were no longer influenced by the 5′ terminal base regardless of the length of the 3′ overhang (Fig. 4d,e, lanes 9–12 and Extended Data Fig. 8h). Thus, R821 has a key role in recognizing the 5′ terminal sequence, disfavouring 5′-G.

To further examine the 5′ terminal base identity, we looked at the sequences of human pre-miRNAs. We found a substantial enrichment of uridine and adenine—bases that are known to facilitate the loading of the Argonaute (AGO) protein[39] (Fig. 4h). Furthermore, we found a marked depletion of guanine compared to cytosine at the 5′ end relative to other neighbouring positions, despite the fact that both bases are equally disfavoured by the AGO protein. This suggests an evolutionary pressure against 5′-G in natural pre-miRNAs.

Finally, by examining miRNAs that are affected by the R821 mutations (Fig. 4f and Extended Data Figs. 8j and 9b), we found that the G base is depleted at the 5′ end of affected miRNAs, in contrast to those that are unaffected (Fig. 4i). This is in line with the structural prediction (Extended Data Fig. 9a) that U, C or A can be inserted into the 5′ pocket, whereas G is incompatible. Thus, owing to the structural restriction posed by R821, hDICER has a nucleotide preference for H (any base but G) at the 5′ terminal position for the recognition of the 5′ end.

Discussion

We have here captured the cryo-EM structure of hDICER–pre-let-7a-1^{GYM} in a cleavage-competent state. Our structure reveals how hDICER can achieve its substrate specificity through extensive contacts (Fig. 5). Most notably, we discover that a frayed 5′ terminal nucleotide other than G ('5′-H') is inserted into the 5′ pocket in the platform domain.

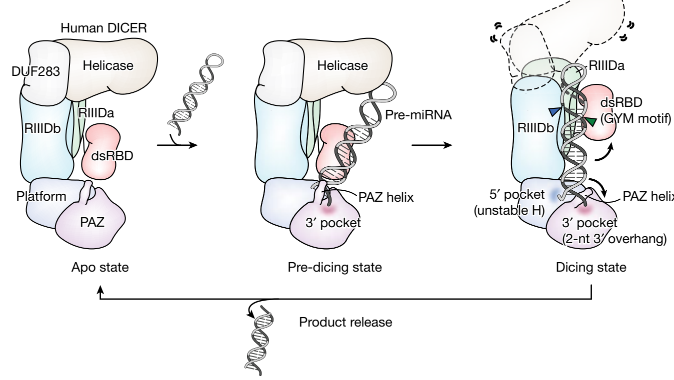


Fig. 5 | Model of the structural transition and substrate recognition of hDICER during the pre-miRNA processing cycle. In the apo state (this study, Protein Data Bank (PDB): 5ZAK)[24], the N-terminal helicase domain, the PAZ helix and the C-terminal dsRBD block the entrance of the pre-miRNA substrate. In the pre-dicing state (PDB: 5ZAL)[24], the 3′ end of the pre-miRNA is stably inserted into the 3′ pocket and the terminal loop contacts the helicase domain while no major conformational changes occur, limiting the access of RNA to the catalytic centre. In transition to the dicing state, three major conformational changes occur, with the helicase domain in highly flexible conformations, the PAZ helix reorienting to allow a simultaneous recognition of the 5′ and 3′ termini, and the dsRBD swinging out to accommodate pre-miRNA in the catalytic centre. After cleavage, the product (miRNA duplex) is released and the enzyme returns to the original closed conformation. Note that the *cis*-elements are recognized by the specific domains of hDICER to allow RNA binding and dictate cleavage sites: the GYM motif by the dsRBD and RIIIDa; the thermodynamically unstable 5′-H by the 5′ pocket; and the 2-nt 3′ overhang by the 3′ pocket.

Thermodynamically unstable 5′-H contributes to efficient and accurate processing. We also find that the dsRBD not only interacts with the RNA backbone but also actively engages in a sequence-specific interaction by recognizing the mismatch of the GYM motif. The dsRBD, through the flexible linker, may scan the upper stem of dsRNA, locate the motif and facilitate the cleavage two nucleotides away from the mismatch.

Combined with the apo and pre-dicing states[24], our dicing-state structure completes the structural landscape of the pre-miRNA processing cycle (see Fig. 5 for a cartoon model and Supplementary Video 3). In the apo state, hDICER exhibits a compact architecture in which the helicase domain, the PAZ helix and the dsRBD limit the access of RNA to the catalytic valley. During the transition to the pre-dicing state, the protein conformation remains largely unchanged, creating only limited protein–RNA contacts through the 3′ pocket, the outer surface of the dsRBD and the helicase domain[24]. Next, the transition from the pre-dicing to the dicing state induces major changes in the helicase domain, the PAZ helix and the dsRBD to fully dock and poise pre-miRNA for cleavage within the catalytic centre. After cleavage, the product is subsequently released from DICER, and transferred to AGO. Further structural investigation will be required to improve our understanding of the AGO-loading step and the role of TRBP in the process.

Comparisons between Dicer homologues provide valuable insights into the mechanism and evolution of eukaryotic small RNA pathways. Some primitive homologues, such as *Giardia* Dicer, do not have the helicase, the PAZ helix or the dsRBD, implying that these domains may have emerged relatively recently for regulatory purposes. The helicase domains of Dicer homologues seem to be particularly diverse even among higher eukaryotes. Our structure shows that the helicase domain of hDICER becomes largely flexible in a dicing state, potentially serving as the regulatory barrier that contributes to substrate specificity. The helicase domain is also known to interact with TRBP, but the mutations or deletions of the helicase domain do not impair or even enhance the dicing activity, suggesting an autoinhibitory

function[14]. Unlike Dicer homologues that processively cleave long dsRNAs using an ATP-dependent translocation by the helicase domain, such as *Drosophila* Dcr-2 and plant Dicer-like proteins[40,41], animal Dicer proteins, the main substrates of which are short hairpin-shaped pre-miRNAs, do not require ATP and cannot perform processive dicing on long dsRNAs[17,19] (Extended Data Fig. 10a). On the basis of sequence similarity, vertebrate Dicer proteins form a separate group from the other Dicer homologues in the miRNA pathway (such as fly Dcr-1), as well as those in the long dsRNA pathway (including fly Dcr-2) (Extended Data Fig. 10b). hDICER prefers hairpins with a flexible terminal loop, but it can also process long dsRNAs—unlike fly Dcr-1, which is highly specialized in pre-miRNA processing[13,18,42,43]. Together with these observations, our current structure suggests that the helicase domain of hDICER may contribute to processing by interacting with a single-stranded RNA region at the initial binding step, but it is not required for the catalytic step and needs to move flexibly to transit to an open conformation. Also supporting this notion, the helicase and DUF283 domains of an oocyte-specific mouse ΔHEL1 Dicer isoform were difficult to model owing to flexibility[44]. Previous structures of *Drosophila* Dcr-1 and its partner Loqs-PB show an overall similar configuration to that of hDICER, but the helicase part seems to be more rigid[45]. Together, these structures suggest conserved yet unique features of vertebrate Dicer proteins that have evolved to act mainly in the miRNA pathway while still maintaining a role in the siRNA pathway.

Collectively, our work on hDICER will provide practical benefits for the rational design of short hairpin RNAs (shRNAs) and Dicer-substrate siRNAs (DsiRNAs) to improve RNAi[46–48]. Moreover, our structure offers a molecular framework for understanding DICER-related diseases in humans. On the basis of the structure, we identified hDICER mutations in the 5′ pocket, which had not been characterized previously. These mutations—R821H and R1003Q—reduce well-known tumour-suppressive miRNAs (for example, let-7 members, miR-324 and miR-142). Given that various DICER mutations have been described in cancers and in a genetic disorder called DICER1 syndrome[49], functional analyses based on the structure will help us to understand how mutations in various parts of hDICER differentially affect the miRNA population and shape the transcriptome in favour of tumorigenesis.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-05723-3.

1. Elbashir, S. M., Lendeckel, W. & Tuschl, T. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.* **15**, 188–200 (2001).
2. Hutvagner, G. et al. A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science* **293**, 834–838 (2001).
3. Lee, Y.-Y., Kim, H. & Kim, V. N. Sequence determinant of small RNA production by DICER. *Nature* https://doi.org/10.1038/s41586-023-05722-4 (2023).
4. Hutvagner, G. & Zamore, P. D. A microRNA in a multiple-turnover RNAi enzyme complex. *Science* **297**, 2056–2060 (2002).
5. Mourelatos, Z. et al. miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev.* **16**, 720–728 (2002).
6. Liu, J. et al. Argonaute2 is the catalytic engine of mammalian RNAi. *Science* **305**, 1437–1441 (2004).
7. Macrae, I. J., Li, F., Zhou, K., Cande, W. Z. & Doudna, J. A. Structure of Dicer and mechanistic implications for RNAi. *Cold Spring Harb. Symp. Quant. Biol.* **71**, 73–80 (2006).
8. MacRae, I. J., Zhou, K. & Doudna, J. A. Structural determinants of RNA recognition and cleavage by Dicer. *Nat. Struct. Mol. Biol.* **14**, 934–940 (2007).
9. Wang, Q. et al. Mechanism of siRNA production by a plant Dicer–RNA complex in dicing-competent conformation. *Science* **374**, 1152–1157 (2021).
10. Sinha, N. K., Iwasa, J., Shen, P. S. & Bass, B. L. Dicer uses distinct modules for recognizing dsRNA termini. *Science* **359**, 329–334 (2018).
11. Su, S. et al. Structural insights into dsRNA processing by *Drosophila* Dicer-2–Loqs-PD. *Nature* **607**, 399–406 (2022).
12. Macrae, I. J. et al. Structural basis for double-stranded RNA processing by Dicer. *Science* **311**, 195–198 (2006).
13. Tsutsumi, A., Kawamata, T., Izumi, N., Seitz, H. & Tomari, Y. Recognition of the pre-miRNA structure by *Drosophila* Dicer-1. *Nat. Struct. Mol. Biol.* **18**, 1153–1158 (2011).
14. Ma, E., MacRae, I. J., Kirsch, J. F. & Doudna, J. A. Autoinhibition of human Dicer by its internal helicase domain. *J. Mol. Biol.* **380**, 237–243 (2008).
15. Chakravarthy, S., Sternberg, S. H., Kellenberger, C. A. & Doudna, J. A. Substrate-specific kinetics of Dicer-catalyzed RNA processing. *J. Mol. Biol.* **404**, 392–402 (2010).
16. Park, J. E. et al. Dicer recognizes the 5′ end of RNA for efficient and accurate processing. *Nature* **475**, 201–205 (2011).
17. Zhang, H., Kolb, F. A., Brondani, V., Billy, E. & Filipowicz, W. Human Dicer preferentially cleaves dsRNAs at their termini without a requirement for ATP. *EMBO J.* **21**, 5875–5885 (2002).
18. Gu, S. et al. The loop position of shRNAs and pre-miRNAs is critical for the accuracy of dicer processing in vivo. *Cell* **151**, 900–911 (2012).
19. Feng, Y., Zhang, X., Graves, P. & Zeng, Y. A comprehensive analysis of precursor microRNA cleavage by human Dicer. *RNA* **18**, 2083–2092 (2012).
20. Tian, Y. et al. A phosphate-binding pocket within the platform–PAZ–connector helix cassette of human Dicer. *Mol. Cell* **53**, 606–616 (2014).
21. Lau, P. W., Potter, C. S., Carragher, B. & MacRae, I. J. Structure of the human Dicer–TRBP complex by electron microscopy. *Structure* **17**, 1326–1332 (2009).
22. Wang, H. W. et al. Structural insights into RNA processing by the human RISC-loading complex. *Nat. Struct. Mol. Biol.* **16**, 1148–1153 (2009).
23. Lau, P. W. et al. The molecular architecture of human Dicer. *Nat. Struct. Mol. Biol.* **19**, 436–440 (2012).
24. Liu, Z. et al. Cryo-EM structure of human Dicer and its complexes with a pre-miRNA substrate. *Cell* **173**, 1549–1550 (2018).
25. Kim, Y. et al. Deletion of human *tarbp2* reveals cellular microRNA targets and cell-cycle function of TRBP. *Cell Rep.* **9**, 1061–1074 (2014).
26. Wilson, R. C. et al. Dicer–TRBP complex formation ensures accurate mammalian microRNA biogenesis. *Mol. Cell* **57**, 397–407 (2015).
27. Provost, P. et al. Ribonuclease activity and RNA binding of recombinant human Dicer. *EMBO J.* **21**, 5864–5874 (2002).
28. Partin, A. C. et al. Cryo-EM structures of human Drosha and DGCR8 in complex with primary microRNA. *Mol. Cell* **78**, 411–422 (2020).
29. Jin, W., Wang, J., Liu, C. P., Wang, H. W. & Xu, R. M. Structural basis for pri-miRNA recognition by Drosha. *Mol. Cell* **78**, 423–433 (2020).
30. Gan, J. et al. Structural insight into the mechanism of double-stranded RNA processing by ribonuclease III. *Cell* **124**, 355–366 (2006).
31. Zhang, H., Kolb, F. A., Jaskiewicz, L., Westhof, E. & Filipowicz, W. Single processing center models for human Dicer and bacterial RNase III. *Cell* **118**, 57–68 (2004).
32. Ma, E., Zhou, K., Kidwell, M. A. & Doudna, J. A. Coordinated activities of human dicer domains in regulatory RNA processing. *J. Mol. Biol.* **422**, 466–476 (2012).
33. Kim, H. et al. Bias-minimized quantification of microRNA reveals widespread alternative processing and 3′ end modification. *Nucleic Acids Res.* **47**, 2630–2640 (2019).
34. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
35. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal* **6**, pl1 (2013).
36. Kim, H. et al. A mechanism for microRNA arm switching regulated by uridylation. *Mol. Cell* **78**, 1224–1236 (2020).
37. Takamizawa, J. et al. Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. *Cancer Res.* **64**, 3753–3756 (2004).
38. Johnson, S. M. et al. RAS is regulated by the let-7 microRNA family. *Cell* **120**, 635–647 (2005).
39. Frank, F., Sonenberg, N. & Nagar, B. Structural basis for 5′-nucleotide base-specific recognition of guide RNA by human AGO2. *Nature* **465**, 818–822 (2010).
40. Zamore, P. D., Tuschl, T., Sharp, P. A. & Bartel, D. P. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101**, 25–33 (2000).
41. Nykanen, A., Haley, B. & Zamore, P. D. ATP requirements and small interfering RNA structure in the RNA interference pathway. *Cell* **107**, 309–321 (2001).
42. Liu, Z., Wang, J., Li, G. & Wang, H. W. Structure of precursor microRNA's terminal loop regulates human Dicer's dicing activity by switching DExH/D domain. *Protein Cell* **6**, 185–193 (2015).
43. Nguyen, T. D., Trinh, T. A., Bao, S. & Nguyen, T. A. Secondary structure RNA elements control the cleavage activity of DICER. *Nat. Commun.* **13**, 2138 (2022).
44. Zapletal, D. et al. Structural and functional basis of mammalian microRNA biogenesis by Dicer. *Mol. Cell* **82**, 4064–4079 (2022).
45. Jouravleva, K. et al. Structural basis of microRNA biogenesis by Dicer-1 and its partner protein Loqs-PB. *Mol. Cell* **82**, 4049–4063 (2022).
46. Paddison, P. J., Caudy, A. A., Bernstein, E., Hannon, G. J. & Conklin, D. S. Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells. *Genes Dev*. **16**, 948–958 (2002).
47. Amarzguioui, M. et al. Rational design and in vitro and in vivo delivery of Dicer substrate siRNA. *Nat. Protoc.* **1**, 508–517 (2006).
48. Kim, D. H. et al. Synthetic dsRNA Dicer substrates enhance RNAi potency and efficacy. *Nat. Biotechnol.* **23**, 222–226 (2005).
49. Hill, D. A. et al. *DICER1* mutations in familial pleuropulmonary blastoma. *Science* **325**, 965 (2009).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2023

Methods

Plasmid construction

The DNA encoding hDICER (RefSeq NM_030621) was PCR-amplified by a human cDNA library and subcloned into either a pX vector (for protein purification) containing an N-terminal His_{10}-eYFP-SUMOstar-Strep or a pCK vector (for ectopic expression in *DICER*-knockout cells) with its original CMV promoter replaced with a PGK promoter. Site-directed mutagenesis was performed to introduce mutations for functional studies.

Cell culture and protein expression and purification

Suspension HEK293E cells were subcultured in a 37 °C shaking incubator with a humidified atmosphere of 8% CO_2, in Dulbecco's modified Eagle's medium (DMEM, WELGENE) supplemented with 5% fetal bovine serum (FBS, WELGENE). For a half-litre culture, 0.15 mg of the plasmid encoding the full-length hDICER, N-terminally tagged with His_{10}-eYFP-SUMOstar-Strep tag, was transiently transfected using 1.5 mg of linear polyethylenimine (PEI) and 1% DMSO. After transfection, the cells were incubated at 33 °C.

The entire protein purification steps were performed at 4 °C. The cells were collected after 72 h, washed with cold PBS and resuspended in buffer A (100 mM Tris-HCl (pH 8.0), 150 mM NaCl, 10% glycerol, 2 mM β-mercaptoethanol and 0.1 mM phenylmethylsulfonyl fluoride) supplemented with EDTA-free Pierce Protease Inhibitor Mini Tablets (Thermo Fisher Scientific), 20 μg ml^{-1} micrococcal nuclease and 5 mM $CaCl_2$. Cells were subsequently sonicated and the lysate was clarified by centrifugation at 35,000*g* for an hour. Clarified lysate was loaded onto a column packed with Ni-NTA Superflow resin (Qiagen) that was pre-equilibrated with buffer A. The resin with the bound protein was washed with 5 column volumes of buffer A supplemented with 40 mM imidazole. The protein was eluted with buffer A supplemented with 200 mM imidazole and incubated with SUMOstar protease (LifeSensors) at 4 °C to cleave the N-terminal tag. The cleavage was confirmed by SDS–PAGE. The protein was loaded onto a column packed with Strep-Tactin Superflow (IBA Lifesciences) resin that was pre-equilibrated with buffer A. The resin was then washed with 5 column volumes of buffer A with 2 mM EDTA and subsequently with 5 column volumes of buffer A without EDTA. The protein was eluted with buffer A with 50 mM biotin. To remove uncleaved fusion protein, the eluate was loaded onto Ni-NTA Superflow resin (Qiagen). The unbound protein was concentrated in a 100-kDa molecular weight cut off Amicon Ultra-15 Centrifugal Filter Unit (Merck). The concentrated protein was subjected to size-exclusion chromatography on a Superose 6 Increase 5/150 GL (GE Healthcare) pre-equilibrated with 50 mM Tris (pH 8.0), 100 mM NaCl and 0.5 mM TCEP. For the apo structure, the protein was concentrated to around 1 mg ml^{-1}, snap-frozen in liquid nitrogen and stored at −80 °C. The same purification method was used for mutant DICER proteins.

In vitro reconstitution of the DICER and pre-miRNA complex

Synthetic pre-let-7a-1^{GYM} (Integrated DNA Technologies) (5′-UGAG GUAGUAGGUUGUAUCGCUUUAGGGUCACACCCACCACUGGGAGAU AGCCAUACAAUCUACUGUCUUUCU-3′) was resuspended in 20 mM Tris (pH 7.5), 80 mM NaCl and 1 mM EDTA. The RNA was heated at 80 °C for 5 min and annealed by slowly decreasing the temperature to 4 °C at a rate of −1 °C min^{-1}. Complex formation was performed by mixing DICER and RNA at final concentrations of 15 μM and 45 μM, respectively, in a buffer containing 50 mM Tris-HCl (pH 8.0), 100 mM NaCl, 0.5 mM TCEP and 3 mM $CaCl_2$. The mixture was incubated on ice for 30 min and loaded onto a Superose 6 Increase 5/150 GL (GE Healthcare) pre-equilibrated with 50 mM Tris (pH 8.0), 100 mM NaCl, 0.5 mM TCEP and 2 mM $CaCl_2$. The fractions containing the complex were pooled, snap-frozen in liquid nitrogen and stored at −80 °C. The concentration of the complex was evaluated on an SDS–PAGE gel and 15% urea-polyacrylamide gel.

Cryo-EM specimen preparation and data collection

Apo-hDICER. An aliquot (2.5 μl) of around 1.6 μM apo-hDICER sample was applied to glow-discharged 300-mesh UltraAuFoil R1.2/1.3 holey-gold grids (Quantifoil), blotted for 2 s with blot force 5 at 15 °C in 100% humidity and vitrified using Vitrobot Mark IV (Thermo Fisher Scientific) at the Center for Macromolecular and Cell Imaging (CMCI) at Seoul National University (SNU). A total of 5,416 movies (1,390 movies at 1.1 Å per pixel and 4,026 movies at 0.87 Å per pixel) were collected in electron-event representation (EER) mode with a defocus range of −1.0 to −2.75 μm on a 200-kV Glacios (Thermo Fisher Scientific, CMCI at SNU) equipped with a Falcon 4 direct electron camera using EPU software (Thermo Fisher Scientific). The movies at 0.87 Å per pixel were then resampled into 1.1 Å per pixel and combined for further processing. Detailed information on data collection is provided in Extended Data Table 1.

hDICER–pre-let-7a-1^{GYM}. An aliquot (2.5 μl) of around 1.0 μM DICER–pre-let-7a-1^{GYM} complex sample was applied to glow-discharged 300-mesh UltraAuFoil R1.2/1.3 holey-gold grids, blotted for 2 s with blot force 5 at 15 °C in 100% humidity and vitrified using Vitrobot Mark IV in SNU CMCI. A total of 5,964 micrographs were collected at a magnification of 105,000× (corresponding to a calibrated pixel size of 0.849 Å) and a defocus range of −0.9 to −2.2 μm on a 300-kV Titan Krios (Thermo Fisher Scientific, Institute of Basic Science, Daejeon, Korea) equipped with a Gatan K3 BioQuantum (Gatan) detector using EPU software (Thermo Fisher Scientific). Detailed information on data collection is provided in Extended Data Table 1.

hDICER–pre-let-7a-1^{GYM} + Mg^{2+}. Around 1.0 μM DICER–pre-let-7a-1^{GYM} complex was incubated with $MgCl_2$ added at a final concentration of 2 mM for more than 10 min at room temperature. The cryo grid was prepared as described above for hDICER–pre-let-7a-1^{GYM}. A total of 2,882 movies at at 1.1 Å per pixel were collected in EER mode with a defocus range of −1.1 to −2.3 μm on a 200-kV Glacios (Thermo Fisher Scientific, CMCI at SNU) equipped with a Falcon 4 direct electron camera using EPU software (Thermo Fisher Scientific).

Data processing and 3D refinement

All image processing was done in cryoSPARC v.3.2 (ref. [50]). Computing resources were used in the CMCI at SNU.

Apo-hDICER. Movies were aligned in 5 × 5 patches in MotionCor2 (ref. [51]), and CTF parameters were estimated with GCTF (ref. [52]). Using template-based autopicking in cryoSPARC, raw particles were initially picked and extracted. After two-dimensional (2D) classification and removing bad particles, 128,635 particles were subjected to 3D heterogeneous refinement using an ab initio model in cryoSPARC. After further 3D classification, global CTF refinements and non-uniform refinement were performed using 128,635 particles, yielding a 4.04-Å map of apo-hDICER based on the gold-standard Fourier shell correlation (FSC) at 0.143. This analysis workflow is illustrated in Extended Data Fig. 2. Detailed information on data processing and 3D refinement is provided in Extended Data Table 1.

hDICER–pre-let-7a-1^{GYM}. Motion correction and CTF estimation was performed in the same process of apo-hDICER. For further automated particle-picking process, an initial particle template was roughly generated based on manually picked 189 particles. Using 3,762,043 particles from an initial template-based picking method, 2D classification and 2D selection was performed. After excluding bad particles and a heterogeneous portion especially containing partial flexible helicase domains, 1,386,301 particles were used for ab initio modelling and 3D classification yielding a homogeneous particle population. Using non-uniform refinement and local refinement, hDICER–pre-let-7a-1 was

reconstructed at a resolution of 3.04 Å map based on the gold-standard FSC at 0.143. This analysis workflow is illustrated in Extended Data Fig. 3. Detailed information on data processing and 3D refinement is summarized in Extended Data Table 1.

hDICER–pre-let-7a-1^{GYM} + Mg^{2+}. Motion correction and CTF estimation was performed using the same process as that for hDICER–pre-let-7a-1^{GYM}. An initial particle template was roughly generated based on manually picked 87 particles, and 2,857,379 particles were automatedly picked by Blob Tuner with a blob diameter of 30 Å to 300 Å. For further automated particle-picking processes, an initial particle template was roughly generated based on the manually picked 189 particles. After excluding bad particles and a heterogeneous portion especially containing partial flexible helicase domains through 2D classification and 2D selection, 279,763 particles were used for ab initio modelling and 3D classification, yielding a homogeneous particle population. Further subclassification was performed by hetero refinement based on 3D templates generated with structural models of apo-like and dicing-state-like DICER in this study.

Model building

Apo-hDICER. Model building started from an initial protein model from the previously reported hDICER–TRBP structure (PDB: 5ZAK)[24]. The initial model was fitted into the density map by Dock in map tool from Phenix v.1.18.1 and manually refined in Coot[53,54]. Then, the model was refined on the Namdinator server using MDFF and Phenix real space refinement default options[55]. After a few rounds, the model was further corrected using Phenix and Coot[56].

hDICER–pre-let-7a-1^{GYM}. An initial protein model from our apo-hDICER structure and an initial pre-let-7a-1 model from a previously reported hDICER structure (PDB: 5ZAL)[24] were rigidly fitted into the density by rigid-body fitting with the Fit in Map tool from Chimera v.1.14. This fitted model was inspected and manually adjusted in Coot, and further refined with phenix.real_space_refine in Phenix, ISOLDE v.1.1.0.

All models were validated by phenix.validation_cryoem[56]. All figures in the manuscripts were illustrated by ChimeraX[57] and PyMol (Schrödinger). For the residual validation analysis, Q-scores for each residue were derived from MapQ of Segger tool[58] plugged in Chimera v.1.15. B-factor values were derived from real space refinement in Phenix ISOLDE v.1.1.0.

In vitro DICER processing assay

All of the RNA oligos described below were chemically synthesized (Integrated DNA Technologies) and gel-purified before use. To prepare pre-miRNA substrates with 1-nt, 2-nt and 3-nt 3′ overhang lengths, RNA oligos (5′-UGAGGUAGUAGGUUGUAUAGUUUUAGGGUCACA CCCACCACUGGGAGAUAACUAUACAAUCUACUGUCUUUC(U)(U)-3′) were radiolabelled at their 5′ ends with [γ-^{32}P]ATP by T4 polynucleotide kinase (Takara). The RNA oligos were then purified using Oligo Clean & Concentrator (Zymo Research) according to the manufacturer's instructions. The eluted RNAs were annealed in a buffer containing 20 mM Tris (pH 7.5), 80 mM NaCl and 1 mM EDTA, by slowly decreasing the temperature to 4 °C at a rate of −1 °C min^{-1}. To prepare pre-miRNA-like duplex RNA substrates, RNA oligos of the 5′-arm with varying 5′ terminal base (5′-(N)GAGGUAGUAGGUUGUAAGUAGAAAGGACAAAGAG-3′) were radiolabelled, purified and annealed to the complementary RNA oligos of the 3′-arm (5′-CUCUUUGUCCAAACUACUUACAACCUACUACCUUAU U(U)-3′ as described above.

The RNA substrates were cleaved with the purified DICER proteins added at a final concentration of 0.5 μM, in a buffer containing 50 mM Tris (pH 7.5), 100 mM NaCl, 0.5 mM TCEP, 2 mM $MgCl_2$ and 5 μg ml^{-1} yeast RNA. After incubation at 37 °C, the reaction was halted by addition of an equivalent volume of 2× RNA Loading Dye (NEB) supplemented with 0.5 mg ml^{-1} proteinase K (Roche). The RNA was then resolved on a 15% urea-polyacrylamide gel, along with 5′-end-radiolabelled synthetic miRNAs and Decade Markers System (Ambion). The product was visualized by phosphorimaging with Typhoon FLA 7000 (GE Healthcare).

DICER rescue experiments and data analysis

DICER-knockout HCT116 cells were cultured in McCoy's 5A medium (WELGENE), supplemented with 10% FBS (WELGENE)[59]. Cells were authenticated by short tandem repeat profiling (ATCC) and confirmed to be mycoplasma-negative. To ectopically express DICER proteins, the cells were transiently transfected with DICER-expressing vectors using FuGENE HD (Promega). The cells were treated with TRIzol (Thermo Fisher Scientific) 24 h or 48 h after transfection to extract total RNAs. For library construction, AQ-seq was performed following the protocol described previously[33], except that 100 μg of total RNAs was used and that NEBNext Ultra II Q5 Master Mix (NEB) was used to amplify the cDNAs. The libraries were sequenced on the NovaSeq 6000 platform.

Data were preprocessed as previously described[33]. In brief, the 3′ adapter with 5′ 4-nt degenerate sequences was removed using cutadapt[60] and FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Next, short, low-quality and artefact reads were filtered out using FASTX-Toolkit. The output reads were mapped to the spike-in reference first and then the unmapped reads were aligned to the human genome (hg38) using BWA[61]. miRNA annotations corresponding to each alignment were retrieved with miRBase release 21 using the intersect tool in BEDTools[62,63]. Results from this data are summarized in Supplementary Table 1.

Statistics and reproducibility

All in vitro assays were repeated at least twice (Figs. 3d and 4d,e and Extended Data Figs. 1b–g, 5e, 7c,d and 8h). The particle distribution pattern of the representative micrograph of apo-DICER (Extended Data Fig. 2b) repetitively appeared in 3,997 macrographs selected from 5,416 micrographs (Extended Data Fig. 2a). The particle distribution pattern of the representative micrograph of DICER–pre-let-7a-1^{GYM} (Extended Data Fig. 3b) repetitively appeared in 4,817 macrographs selected from 5,974 micrographs (Extended Data Fig. 3a).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The structural models and density maps have been deposited in the PDB under the accession codes 7XW3 (apo-hDICER) and 7XW2 (hDICER–pre-let-7a-1^{GYM}), as listed in Extended Data Table 1. The raw images have been deposited in the Electron Microscopy Data Bank (EMDB) under the accession codes EMD-33490 (apo-hDICER) and EMD-33489 (hDICER–pre-let-7a-1^{GYM}), as listed in Extended Data Table 1. Other structural models cited in this study for analysis (5ZAL, 5ZAK, 2EZ6, 7VG2, 7VG3, 4NGD, 4NHA and 4NH6) are also accessible through the PDB. The rescue data were deposited to the Gene Expression Omnibus (GSE215867).

Code availability

Custom analysis codes are available at https://github.com/haedong-kim615/dicer_dicing_state_structure.

50. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
51. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
52. Zhang, K. Gtcf: real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
53. Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr. D* **75**, 861–877 (2019).

54. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
55. Kidmose, R. T. et al. Namdinator—automatic molecular dynamics flexible fitting of structural models into cryo-EM and crystallography experimental maps. *IUCrJ* **6**, 526–531 (2019).
56. Afonine, P. V. et al. Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. D* **74**, 531–544 (2018).
57. Goddard, T. D. et al. UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).
58. Pintilie, G. et al. Measurement of atom resolvability in cryo-EM maps with Q-scores. *Nat. Methods* **17**, 328–334 (2020).
59. Kim, Y. K., Kim, B. & Kim, V. N. Re-evaluation of the roles of DROSHA, Export in 5, and DICER in microRNA biogenesis. *Proc. Natl Acad. Sci. USA* **113**, E1881–1889 (2016).
60. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.j.* **17**, 10–12 (2011).
61. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
62. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42**, D68–73 (2014).
63. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

Acknowledgements We thank J.-S. Woo for the mammalian cell transfection protocol and Y.-G. Choi, S.-M. Ji, J. Yang, D.-E. Choi, S. Bang and E. Kim for technical assistance. This research was supported by the Institute for Basic Science from the Ministry of Science and ICT of Korea (IBS-R008-D1 to Y.-Y.L., H.K. and V.N.K.); BK21 research fellowships from the Ministry of Education of Korea (to Y.-Y.L. and H.K.); and the National Research Foundation of Korea (NRF-2018-Global Ph.D. Fellowship Program to Y.-Y.L. and NRF-2015-Global Ph.D. Fellowship Program to H.K). S.-H.R. acknowledges financial support from the Creative-Pioneering Researchers Program through Seoul National University, NRF grants (2019M3E5D6063871, 2019R1C1C1004598, 2020R1A5A1018081 and 2021M3A9I4021220) and the SUHF Foundation. Computing resources were used in the CMCI at SNU and the Global Science Experimental Data Hub Center (GSDC) at Korea Institute of Science and Technology Information (KISTI).

Author contributions All of the authors conceived the project. V.N.K. and S.-H.R. collected financial support. Y.-Y.L. performed protein purification and cryo-EM sample preparation. Y.-Y.L. and H.K. performed biochemical and cellular experiments. H.L. and S.-H.R. performed structural studies. H.K. performed bioinformatic analyses.

Competing interests The authors declare no competing interests.

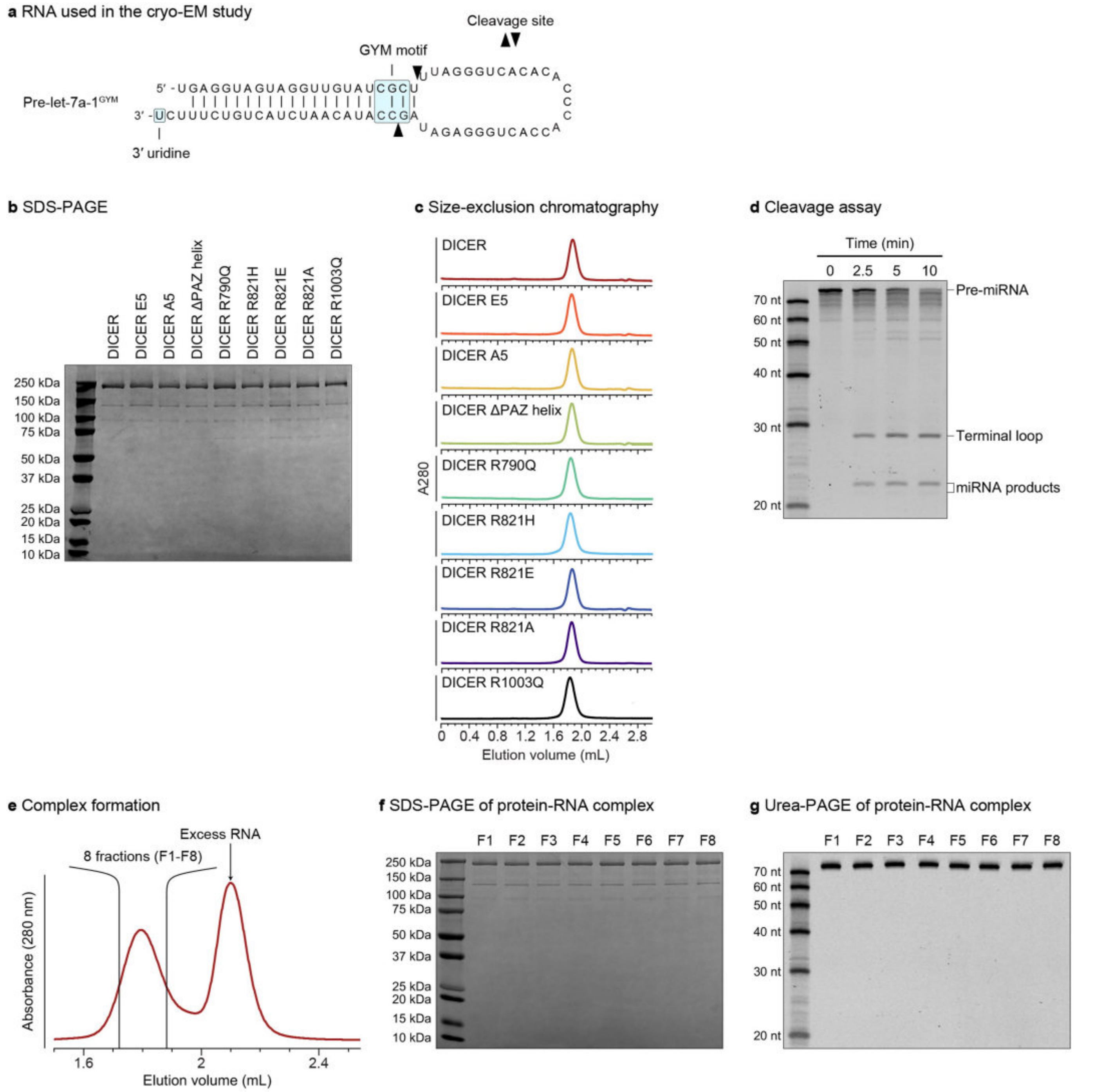
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-05723-3.

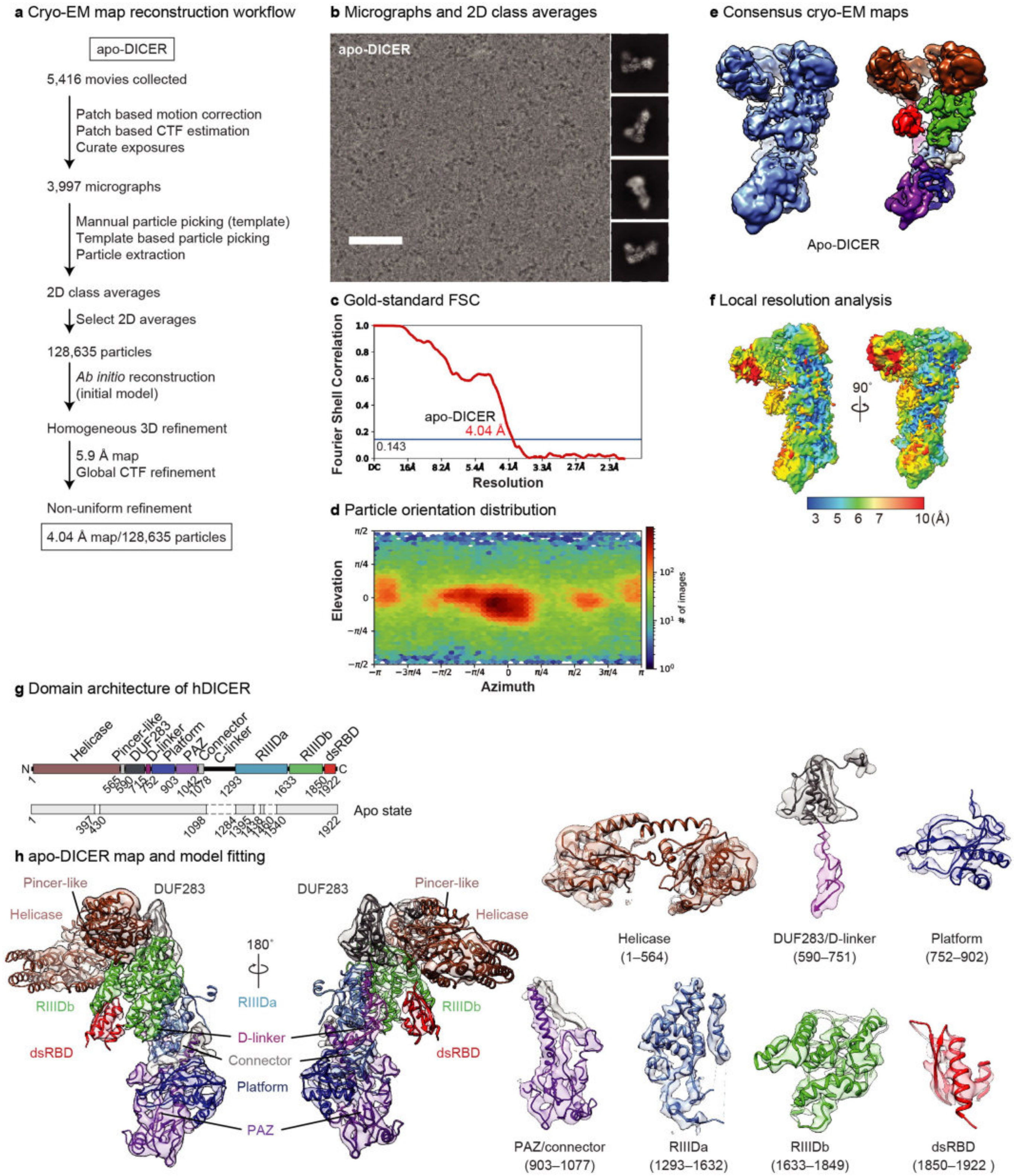
Correspondence and requests for materials should be addressed to V. Narry Kim or Soung-Hun Roh.

Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

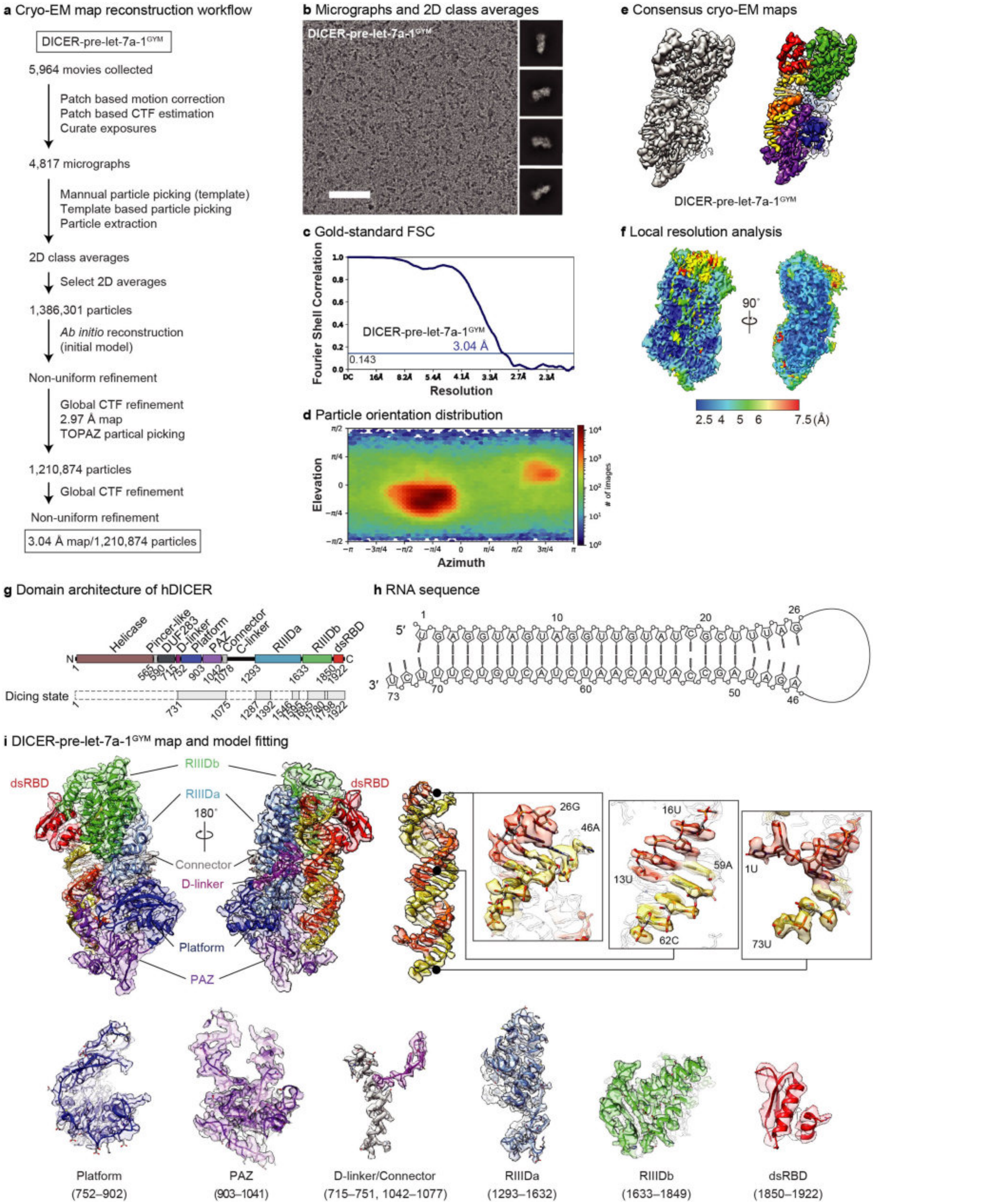
Reprints and permissions information is available at http://www.nature.com/reprints.



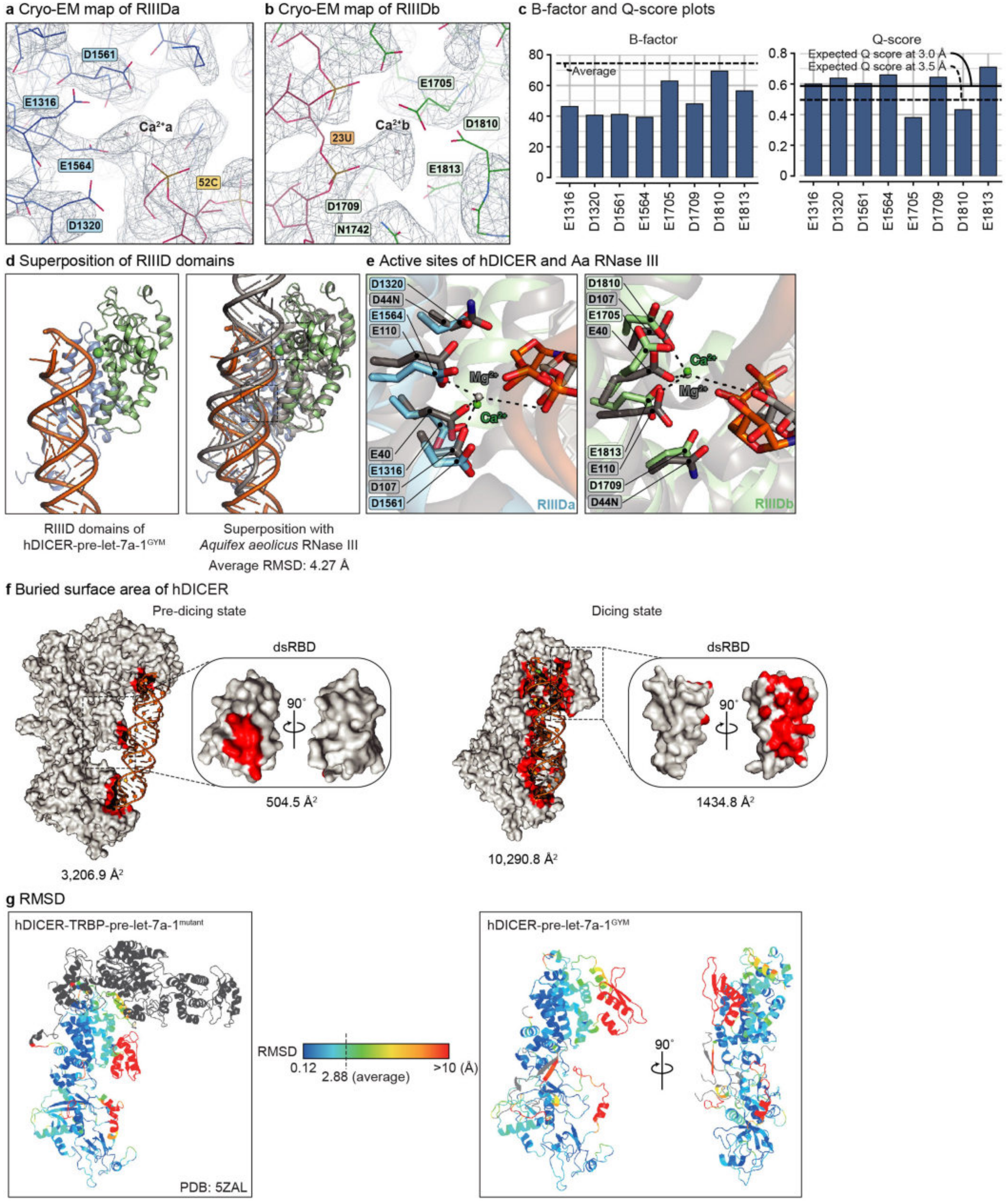
Extended Data Fig. 1 | Purification of hDICER and the hDICER–RNA complex. a, The sequence of pre-let-7a-1^{GYM} used for structural determination. **b**, SDS–PAGE of wild-type and mutant hDICER proteins. **c**, Size-exclusion chromatography of purified proteins. **d**, In vitro processing of pre-let-7a-1 by purified hDICER. **e**, Size-exclusion chromatography of the hDICER–pre-let-7a-1^{GYM} complex. **f**, SDS–PAGE of the hDICER–pre-let-7a-1^{GYM} complex visualized by Coomassie blue staining. Protein concentration for each fraction was estimated by Bradford protein assay, and the same amount of protein was loaded for each fraction. **g**, Urea-PAGE of the hDICER–pre-let-7a-1^{GYM} complex visualized by SYBR gold staining. RNA concentration for each fraction was estimated by absorbance at 260 nm, and the same amount of RNA was loaded for each fraction. For gel source data, see Supplementary Fig. 1.



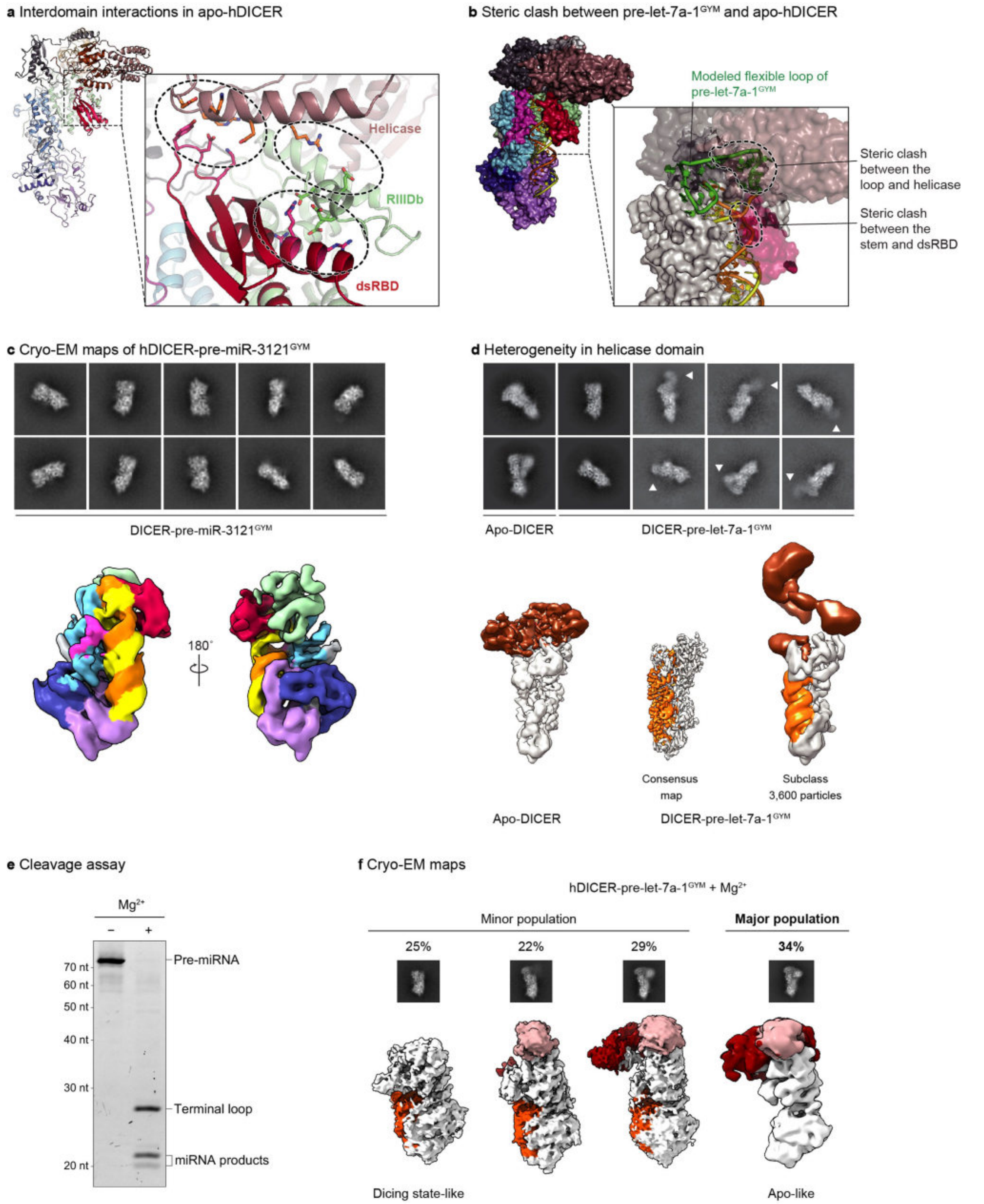
Extended Data Fig. 2 | Cryo-EM image processing procedure for apo-hDICER. **a**, Overview of the image processing procedure (see Methods). **b**, Representative micrograph and 2D class averages of the apo-hDICER (scale bar, 50 nm). **c**, Gold-standard FSC at 0.143 of the apo-hDICER. **d**, Angular particle distribution heat map. **e**, Consensus map of apo-hDICER. Each domain is indicated in a different colour. **f**, Local-resolution analysis shown in rainbow scale. **g**, Domain organization of hDICER with colour code for each domain. Schematics for the apo state shows amino acid residues included (solid lines) or not included (dashed lines) in the model. **h**, Atomic model fitting to the map of apo-hDICER.



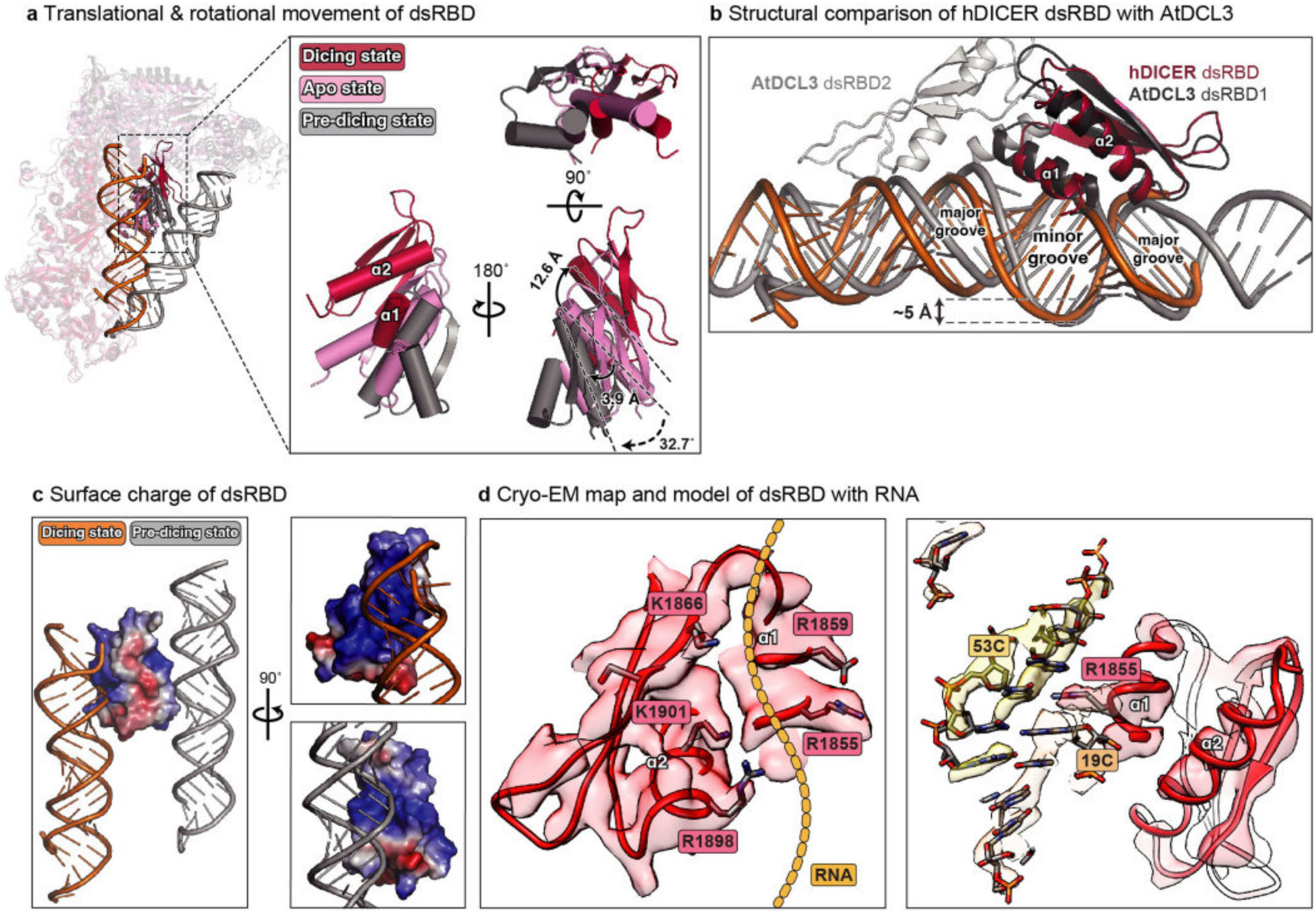
Extended Data Fig. 3 | Cryo-EM image processing procedure for hDICER–pre-let-7a-1GYM. **a**, Overview of the image processing procedure (see Methods). **b**, Representative micrograph and 2D class averages of hDICER–pre-let-7a-1GYM (scale bar, 50 nm). **c**, Gold-standard FSC at 0.143 of the hDICER–pre-let-7a-1GYM. **d**, Angular particle distribution heat map. **e**, Consensus map of hDICER–pre-let-7a-1GYM. Each domain is indicated in a different colour. **f**, Local-resolution analysis shown in rainbow scale. **g**, Domain organization of hDICER with colour code for each domain. Schematics for the dicing state shows amino acid residues included (solid lines) or not included (dashed lines) in the model. **h**, Sequence of pre-let-7a-1GYM in the model. **i**, Atomic model fitting to the map of hDICER–pre-let-7a-1GYM.



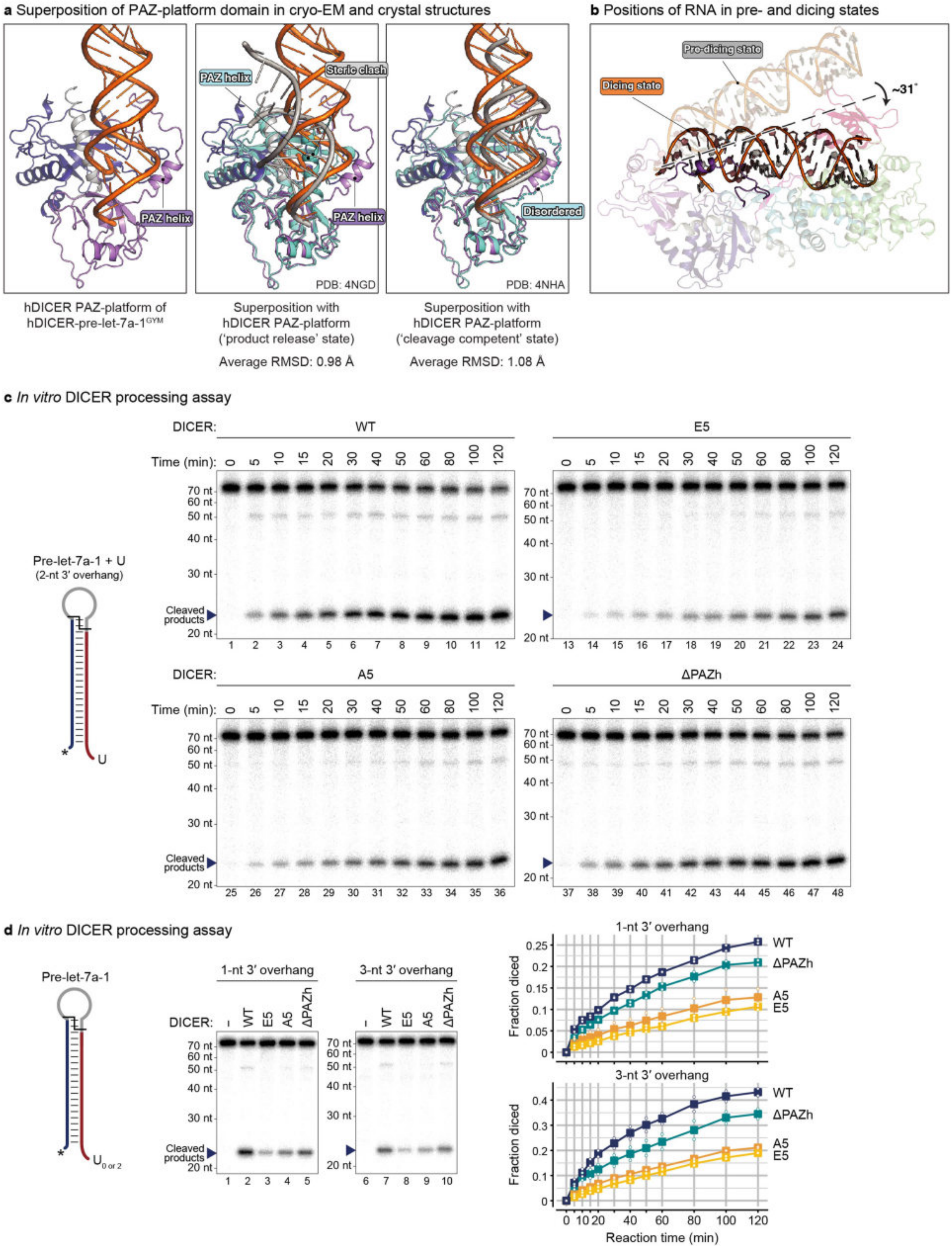
Extended Data Fig. 4 | Overall structure of hDICER in a dicing state.
a, Cryo-EM map of the catalytic site created by RIIIDa. **b**, Cryo-EM map of the catalytic site created by RIIIDb. **c**, B-factor and Q-score plots for active site residues in the hDICER–let-7a-1^{GYM} complex structure. Q-scores for each residue were derived from MapQ of Segger tool plugged in Chimera v.1.15. B-factor values were derived from real space refinement in Phenix ISOLDE v.1.1.0. **d**, Superposition of RIIID domains of hDICER (this study) and Aa RNase III (PDB: 2EZ6, grey)[30]. **e**, Active sites of hDICER and Aa RNase III (PDB:2EZ6, grey)[30]. **f**, Buried surface area of hDICER in a pre-dicing state (PDB: 5ZAL)[24] and a dicing state. **g**, RMSD of hDICER–pre-let-7a-1^{GYM} (this study) compared to hDICER–TRBP-pre-let-7a-1^{mutant} (PDB: 5ZAL)[24]. Residues not resolved in the dicing state are coloured in grey.



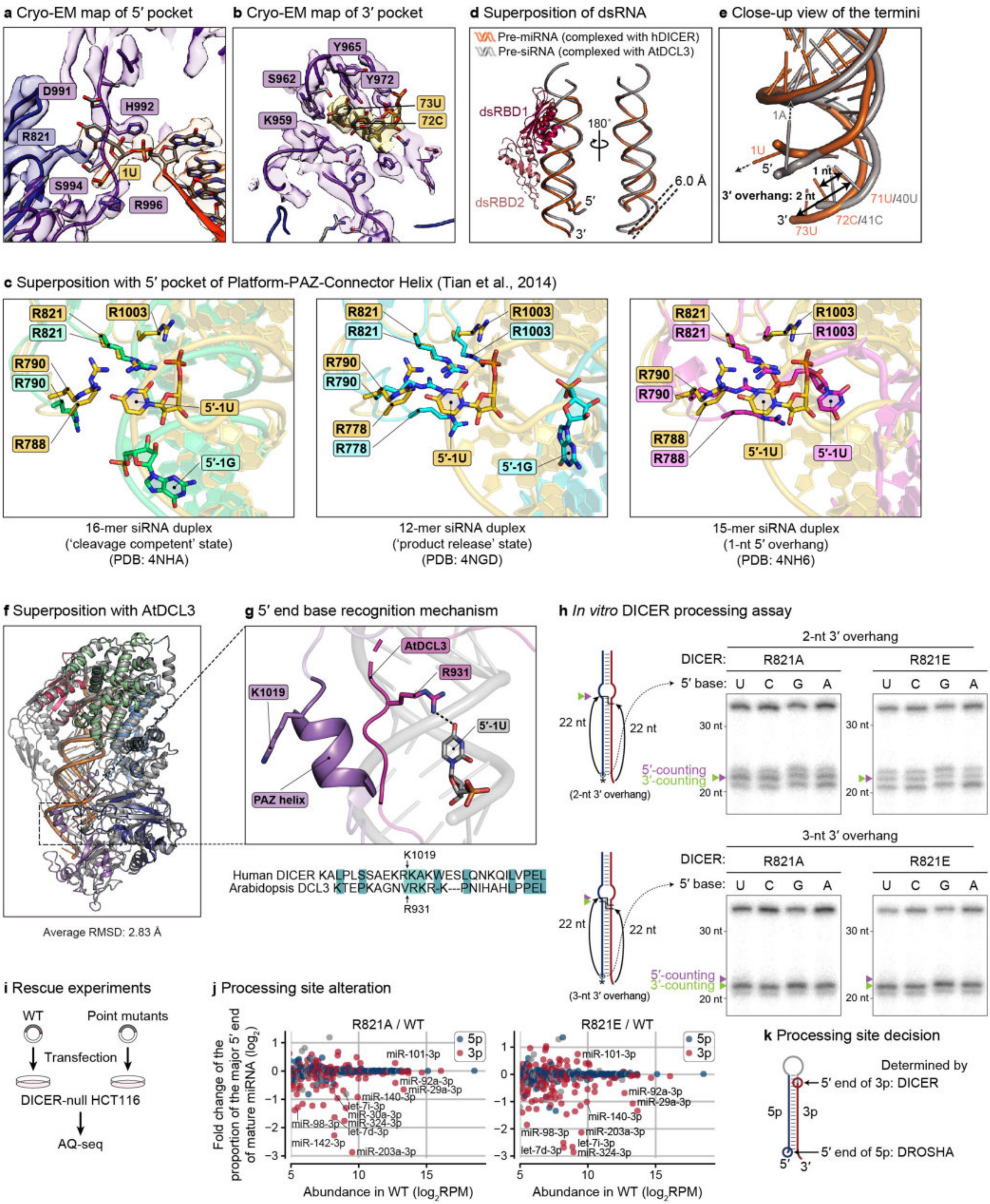
Extended Data Fig. 5 | The structure of the helicase domain. **a**, Interdomain interactions in apo-hDICER. **b**, Steric clash between pre-let-7a-1^{GYM} and apo-hDICER. **c**, Cryo-EM map of the hDICER–pre-miR-3121^{GYM} complex in a dicing state. **d**, Selected 2D class averages and 3D maps showing heterogeneity in the helicase domain. White arrowhead indicates the location of the helicase domain in 2D averages. Bound RNA density is indicated in orange. **e**, Urea-PAGE of hDICER–pre-let-7a-1^{GYM} complex incubated with or without $MgCl_2$ for 10 min at room temperature, visualized by SYBR gold staining. For gel source data, see Supplementary Fig. 1. **f**, Selected 2D class averages and 3D maps showing heterogeneity of the helicase domain of hDICER–pre-let-7a-1^{GYM} complex.



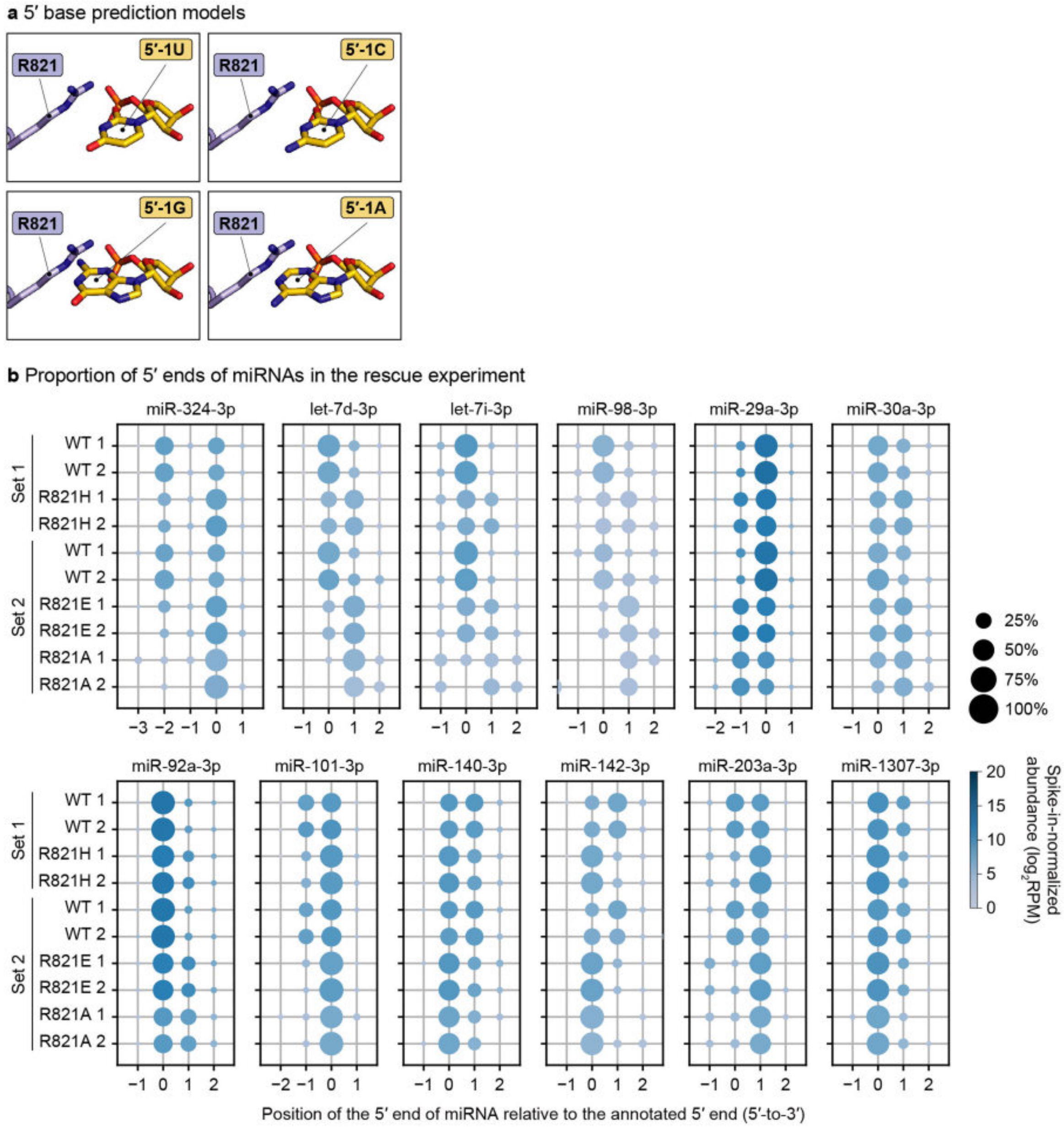
Extended Data Fig. 6 | The structure of dsRBD in different states. **a**, Conformational changes of the dsRBD in the apo (this study), dicing (this study) and pre-dicing states (PDB: 5ZAL)[24]. **b**, Superposition of the dsRBDs of hDICER and AtDCL3 (PDB: 7VG2)[9]. **c**, Surface charge of the dsRBD, with the dsRNA–dsRBD interface in dicing and pre-dicing states (PDB: 5ZAL)[24]. **d**, Cryo-EM map and model of the hDICER dsRBD with dsRNA.



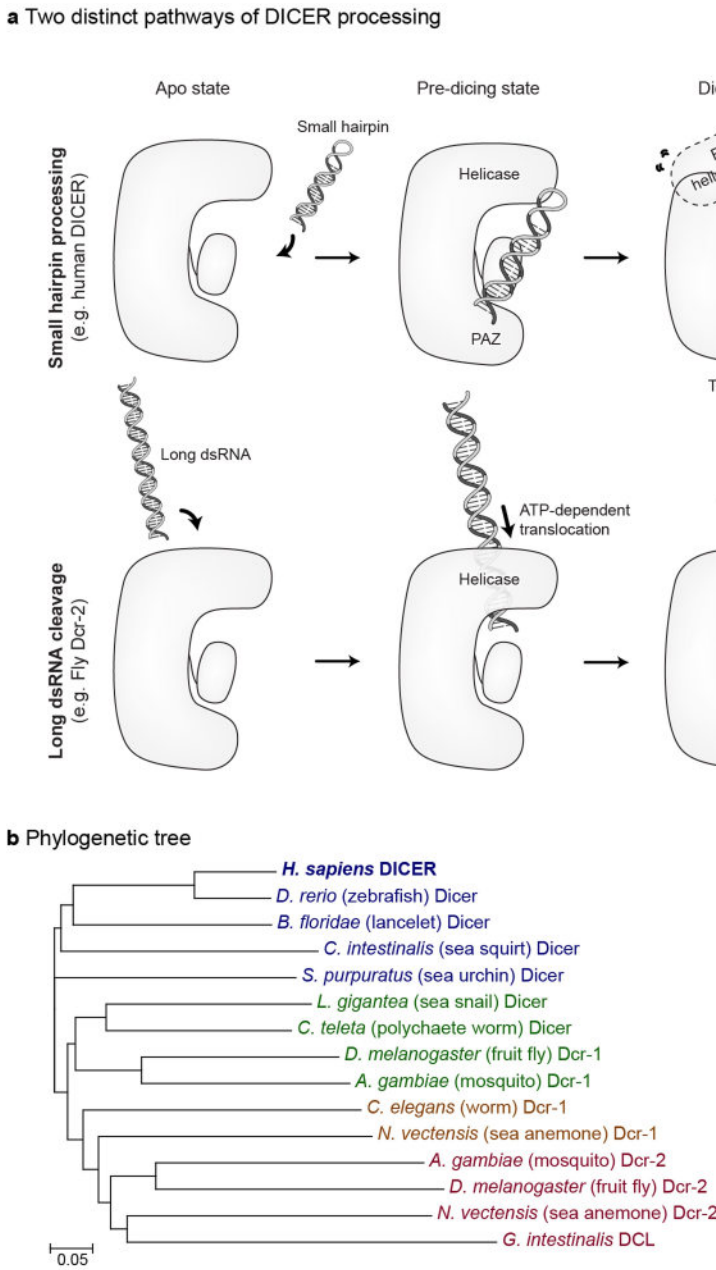
Extended Data Fig. 7 | The PAZ helix rearranges to interact with pre-miRNA in a cleavage-competent position. a, Superposition of hDICER PAZ–platform domain in the cryo-EM structure (this study) and in the crystal structure (PDB: 4NHA, grey)[20]. **b**, Changes in the position of the pre-miRNA in a dicing state (this study) and a pre-dicing state (PDB: 5ZAL)[24]. **c**, In vitro processing of pre-let-7a-1 with a 2-nt 3′ overhang. *, radiolabelled 5′ phosphate. **d**, In vitro processing of pre-let-7a-1 with a 1-nt 3′ overhang (lanes 1–5) or a 3-nt 3′ overhang (lanes 6–10). Relative cleavage was calculated by quantifying the band intensity (1 − uncleaved/input). Squares indicate mean ($n = 2$, independent experiments). *, radiolabelled 5′ phosphate. For gel source data, see Supplementary Fig. 1.



Extended Data Fig. 8 | End recognition mechanism of hDICER. a, Cryo-EM map of the 5′ pocket. **b**, Cryo-EM map of the 3′ pocket. **c**, Superposition of hDICER–pre-let-7a-1^{GYM} and Platform–PAZ–Connector Helix (PDB: 4NHA, 4NGD and 4NH6)[20]. **d**, Superposition of dsRNAs complexed with hDICER and AtDCL3 (PDB: 7VG2)[9]. **e**, Close-up view of the ends of the dsRNAs complexed with hDICER and AtDCL3 (PDB: 7VG2, grey)[9]. **f**, Superposition of hDICER in a dicing state and AtDCL3-pre-siRNA complex (PDB: 7VG3, magenta)[9]. **g**, 5′ terminal base recognition by AtDCL3 via PAZ region corresponding to the PAZ helix of hDICER. **h**, In vitro DICER processing of pre-miRNA-like duplex with a 2-nt or 3-nt 3′ overhang. The base opposite to the varying sequence is A on the 3p strand. For gel source data, see Supplementary Fig. 1. *, radiolabelled 5′ phosphate. **i**, Schematic outline of the rescue experiment ($n = 2$, biological replicates). **j**, Changes in cleavage accuracy, estimated with the fold change of the proportion of the major 5′-isomiR. For a given miRNA, the most abundant 5′-isomiR was identified in the wild-type sample. Grey, unannotated strand. **k**, DROSHA/DICER cleavage sites dictated by 5′ ends of mature miRNAs.



Extended Data Fig. 9 | Rescue experiments with DICER 5′ pocket mutants. **a**, Predicted structural effect of the 5′ end base substitutions on the interaction with the 5′ pocket. **b**, Examples of altered processing sites observed in the rescue experiments. Note that the DICER cleavage sites can be inferred from the 5′ end of 3p miRNAs. miRNA isoforms beginning at the indicated position are plotted with circles, with the size of the circle reflecting the proportion of the cleavage-site usage at the given position.



Extended Data Fig. 10 | Distinct functions and evolution of Dicer proteins in two small RNA pathways. a, Comparison of the substrate RNA movement during DICER processing between two small RNA pathways. In the miRNA pathway, a hairpin-shaped small RNA (pre-miRNA) is bound to DICER by the helicase and PAZ domains. For cleavage, the helicase domain becomes flexible to accommodate the pre-miRNA into the catalytic centre. By contrast, in the siRNA pathway, a long dsRNA comes into DICER by passing through the helicase domain. The ATP-dependent translocation by the helicase domain leads to processive cleavage of long dsRNAs. **b**, A phylogenetic tree of Dicer homologues. The scale bar indicates the length for the indicated frequency of amino acid variation.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	#1 apo-hDICER (EMDB-33490) (PDB 7XW3)	#2 hDICER-let7a-1^{GYM} (EMDB-33489) (PDB 7XW2)
Data collection and processing		
Magnification	120,000	105,000
Voltage (kV)	200	300
Electron exposure (e–/Å^2)	44.163	40
Defocus range (μm)	-1.0 ~ -2.75	-0.9 ~ -2.2
Pixel size (Å)	1.102	0.849
Symmetry imposed	C1	C1
Initial particle images (no.)	189,908	1,386,301
Final particle images (no.)	128,635	1,210,874
Map resolution (Å)	4.04	3.04
FSC threshold	0.143	0.143
Map resolution range (Å)	3.0-10.0	2.5-7.5
Refinement		
Initial model used (PDB code)	5ZAK	7XW3
Model resolution (Å)	4.04	4.03
FSC threshold	0.143	0.143
Model resolution range (Å)	3.0-10.0	2.5-7.5
Map sharpening *B* factor (Å^2)	-120.8	-157.2
Model composition		
Non-hydrogen atoms	12087	6997
Protein residues	1502	725
Nucleotides	-	54 (RNA)
Ligands	--	2 (Ca)
B factors (Å^2)		
Protein	98.43	74.38
Nucleotides	--	83.33 (RNA)
Ligands	--	30.00 (Ca)
R.m.s. deviations		
Bond lengths (Å)	0.009	0.007
Bond angles (°)	1.238	1.058
Validation		
MolProbity score	2.49	1.77
Clashscore	15.50	7.50
Poor rotamers (%)	2.44	0.92
Ramachandran plot		
Favored (%)	91.54	95.10
Allowed (%)	8.39	4.90
Disallowed (%)	0.07	0.00

The top part of the table provides conditions and parameters for the data collection process. The bottom part contains statistic values related to the refinement of density maps and molecular models.

nature portfolio

Corresponding author(s): V. Narry Kim, Soung-Hun Roh

Last updated by author(s): Nov 28, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Cryo-EM data were collected using EPU software (version 2.10.0.5REL). Preprocessing for the sequencing data was done with cutadapt 1.8.3, FASTX Toolkit 0.0.13.2, bwa 0.7.10-r789, bedtools v2.25.0.
Data analysis	Cryo-EM data analysis was conducted using cryoSPARC (version 3.2), MotionCor2 (version 1.3.0) and GCTF (version 1.06). Initial model fitting was carried out using chimera (verson 1.14) and Molecular Dynamics Flexible Fitting (MDFF) provided by Namdinator server (www.namdinator.au.dk). Further model building and refinement process for cryoEM structures were performed using COOT (Version 0.9) and ISOLDE (version 1.1.0) included in PHENIX (version 1.18.1). Visualization of structural models were performed using Chimera X (version l.3rc202111060828) and PyMol (version 2.4.1). Custom analysis codes are available at https://github.com/haedongkim615/dicer_dicing_state_structure.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The cryo-EM structures of the apo-hDICER and hDICER-pre-let7a-1-GYM have been deposited to Protein Data Bank (www.pdb.org) with PDB access codes of 7XW3 and 7XW2, respectively. The cryo-EM density maps of apo-hDICER and hDICER-pre-let7a-1-GYM have been deposited to EMDB under the access codes of EMD-33490 and EMD-33489, respectively. There are no restrictions on data availability. For material requests, please contact the corresponding authors.
Other structural models cited in this study for analysis (5ZAL, 5ZAK, 2EZ6, 7VG2, 7VG3, 4NGD, 4NHA, and 4NH6) are also accessible on PDB. The rescue data were deposited to the GEO repository (accession number: GSE215867).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Not applicable
Population characteristics	Not applicable
Recruitment	Not applicable
Ethics oversight	Not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Described in the corresponding figure legends. No sample-size calculation was performed. For in vitro processing assays, we have performed independent experiments, which is sufficient to estimate the variation.
Data exclusions	No data was excluded.
Replication	Experiments were done with replicates as indicated in the figure legends. Independent methods, such as in vitro DICER processing, were in agreement.
Randomization	This project does not involve allocation into experimental groups.
Blinding	Blinding was not relevant to this study. The results from in vitro processing assays are visualized by urea-PAGE as individual bands that are quantified using software (i.e., ImageJ), requiring no subject interpretation or judgement.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	HCT116 cells
Authentication	Authenticated using the ATCC short tandem repeat profiling. Last performed on Mar 25th, 2022.
Mycoplasma contamination	Cell lines were confirmed mycoplasma negative prior to use.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used in this study.

Article

H3K4me3 regulates RNA polymerase II promoter-proximal pause-release

https://doi.org/10.1038/s41586-023-05780-8

Received: 7 December 2021

Accepted: 2 February 2023

Published online: 1 March 2023

Open access

Check for updates

Hua Wang[1,2], Zheng Fan[3,4,5], Pavel V. Shliaha[6], Matthew Miele[6], Ronald C. Hendrickson[6], Xuejun Jiang[1,2] & Kristian Helin[1,2,3,4,5] ✉

Trimethylation of histone H3 lysine 4 (H3K4me3) is associated with transcriptional start sites and has been proposed to regulate transcription initiation[1,2]. However, redundant functions of the H3K4 SET1/COMPASS methyltransferase complexes complicate the elucidation of the specific role of H3K4me3 in transcriptional regulation[3,4]. Here, using mouse embryonic stem cells as a model system, we show that acute ablation of shared subunits of the SET1/COMPASS complexes leads to a complete loss of all H3K4 methylation. Turnover of H3K4me3 occurs more rapidly than that of H3K4me1 and H3K4me2 and is dependent on KDM5 demethylases. Notably, acute loss of H3K4me3 does not have detectable effects on transcriptional initiation but leads to a widespread decrease in transcriptional output, an increase in RNA polymerase II (RNAPII) pausing and slower elongation. We show that H3K4me3 is required for the recruitment of the integrator complex subunit 11 (INTS11), which is essential for the eviction of paused RNAPII and transcriptional elongation. Thus, our study demonstrates a distinct role for H3K4me3 in transcriptional pause-release and elongation rather than transcriptional initiation.

英文杂志全球首发QQ群:855583774

Histone modifications are closely linked to transcription regulation and are involved in processes that determine cell fate, development and disease[5,6]. Methylation of H3 lysine 4 (H3K4) is one of the most studied modifications owing to its association with gene expression and cancer[2,7]. H3K4 methylation is deposited by SET1/COMPASS complexes that contain different lysine methyltransferases and several essential subunits, including the six catalytic subunits SETD1A, SETD1B and MLL1–4. H3K4me3 is enriched at transcription start sites (TSSs) and is believed to promote transcription through the recruitment of PHD-domain-containing proteins involved in transcription initiation, such as TATA-box-binding protein associated factor 3 (TAF3)[1,8]. Moreover, H3K4me3 has been reported to counteract DNA methylation[9] and repressive histone modifications such as H3K9me3 and H3K27me3[10,11]. Furthermore, the broad H3K4me3 domains, prominently found in preimplantation embryos[10] and somatic cells[12], have been proposed to ensure transcriptional consistency of essential genes to maintain cell identity.

Previous studies have addressed the roles of various components of the SET1/COMPASS complexes, such as CFP1[13], WDR5[14], DPY30[15], MLL1/2[16], MLL3/4[17] and SETD1A/B[18], in mammalian cells by inhibiting their expression or deleting their respective genes. However, observations of the effects on gene expression in these studies have not produced uniform results. These inconsistencies may in part be due to redundant functions of the components of the SET1/COMPASS complexes. Moreover, inadequate temporal resolution of previous knockdown or knockout studies has rendered it difficult to establish a direct role for H3K4 methylation in gene expression[19]. In this study, we determined the transcriptional effects of acute loss of H3K4me3 by targeted degradation of core components of the SET1/COMPASS complexes. Our data show that H3K4me3 has a key role in regulating RNAPII pause-release; however, notably, we did not detect a role for H3K4me3 in regulating transcription initiation.

Models to study H3K4 methylation

To study the role of H3K4me3 in transcription, we developed model systems for the acute depletion of either DPY30 or RBBP5, two core components that have been reported to be integral and shared components of all the SET1/COMPASS family of H3K4 methyltransferase complexes, in mouse embryonic stem (mES) cell lines (Fig. 1a and Extended Data Fig. 1a,b). To do this, we generated isogenic mES cells expressing DPY30 fused to the miniAID degron tag[20] (DPY30–mAID) or RBBP5 fused to the FKBP12(F36V) degron tag[21] (RBBP5–FKBP) (Extended Data Fig. 1c). Auxin treatment led to undetectable levels of DPY30 within 1 h, and treatment with auxin for 24 h caused a substantial decrease in mono-, di- and trimethylation of H3K4 (Fig. 1b and Extended Data Fig. 1d). Similarly, treatment of RBBP5–FKBP cells with dTAG-13 also caused acute loss of H3K4 methylation after the depletion of RBBP5 (Fig. 1c). Thus, these results show that targeting DPY30 or RBBP5 of the SET1/COMPASS complex leads to a rapid loss of H3K4 methylation.

Notably, H3K4me3 levels were strongly decreased within 2 h of DPY30- and RBBP5-induced degradation, whereas a substantial decrease in H3K4me1 and H3K4me2 levels was achieved only after 24 h (Fig. 1b,c). The binding of DPY30 and RBBP5 was completely lost genome-wide, as determined by chromatin immunoprecipitation with sequencing (ChIP–seq) analysis after treatment for 2 h with auxin or dTAG-13, respectively.

[1]Cell Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [2]Center for Epigenetics Research, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [3]The Institute of Cancer Research, London, United Kingdom. [4]Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Copenhagen, Denmark. [5]The Novo Nordisk Foundation Center for Stem Cell Biology (Danstem), University of Copenhagen, Copenhagen, Denmark. [6]Microchemistry and Proteomics Core Facility, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ✉e-mail: kristian.helin@icr.ac.uk

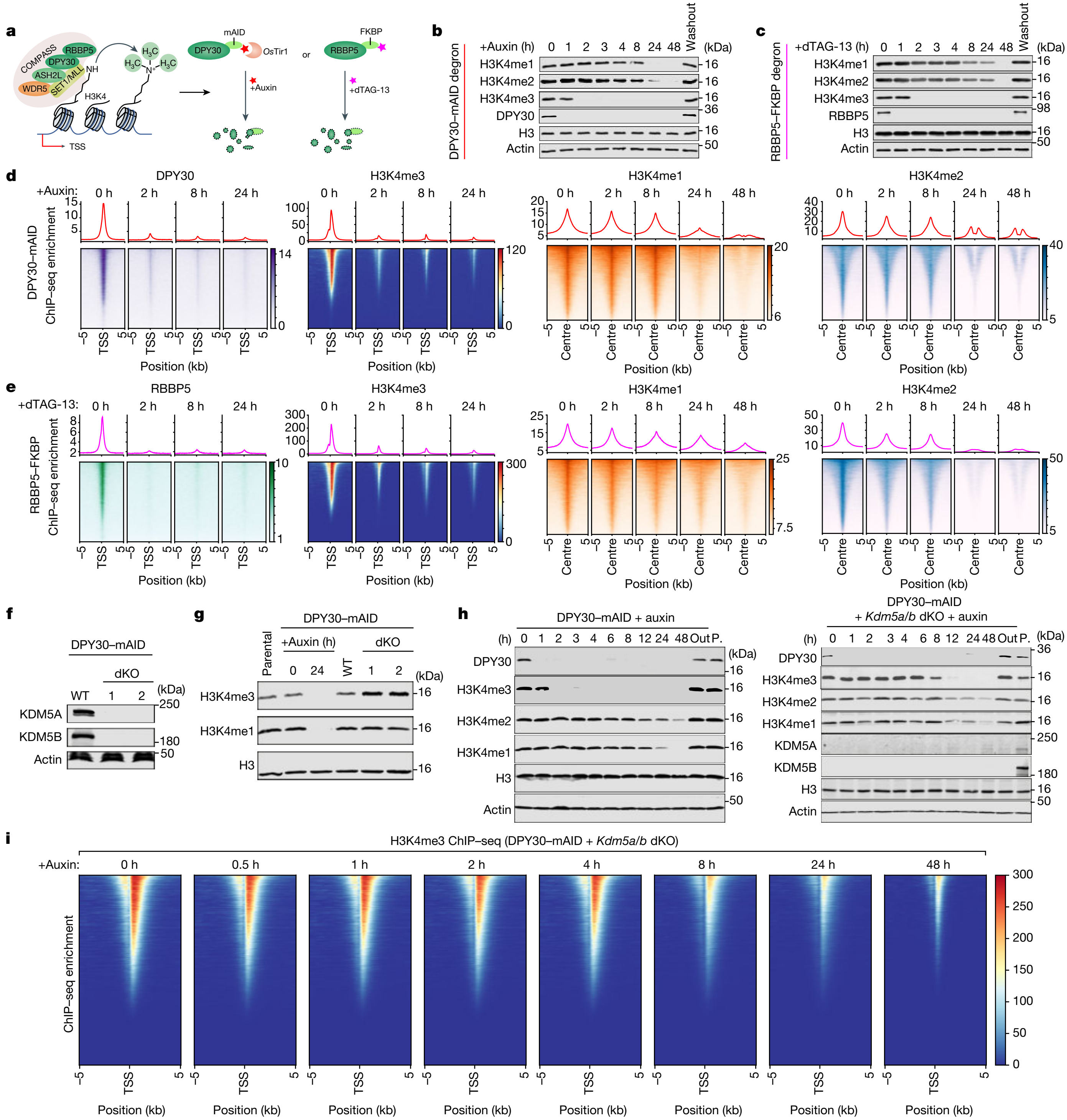


Fig. 1 | Acute depletion of SET1/COMPASS core subunits reveals rapid turnover of H3K4me3. a, Schematic of the degron systems for the targeted degradation of DPY30 and RBBP5. **b**,**c**, Immunoblot analysis of DPY30, RBBP5 and H3K4me1–3 levels at the indicated times after treatment with 500 nM auxin (**b**) or 500 nM dTAG-13 (**c**). Washout, degron ligand was washed out for 48 h. **d**,**e**, ChIP–seq heat maps and profiles were generated from control and auxin-treated DPY30–mAID cells (**d**) and dTAG-13-treated RBBP5–FKBP cells (**e**). For DPY30, RBBP5 and H3K4me3 ChIP–seq, the signal was plotted over the TSSs (TSS ± 5 kb) of protein-coding genes. For H3K4me1 and H3K4me2 ChIP–seq, the signal was plotted over their centre peaks (peak centre ± 5 kb), which are called from steady-state mES cells. Sites were sorted by the ChIP–seq signals at 0 h. **f**, Immunoblot analysis of KDM5A and KDM5B in DPY30–mAID cells and two independently isolated dKO cell lines. β-Actin was used as the loading control. **g**, Immunoblot analysis of H3K4me3 and H3K4me1 levels in DPY30–mAID, control and *Kdm5a/b*-dKO cells. Histone H3 was used as the loading control. **h**, Immunoblot analysis of DPY30, H3K4me1–3, KDM5A and KDM5B at the indicated times after auxin treatment. Out, degron ligand was washed out for 48 h; P, parental cells. **i**, H3K4me3 ChIP–seq heat maps in DPY30–mAID *Kdm5a/b*-dKO cells. The signal was plotted over the TSSs (TSS ± 5 kb) of protein-coding genes. Rows are sorted by decreasing ChIP–seq occupancy in the auxin 0 h cells.

Moreover, all three methylated forms of H3K4 were strongly decreased genome-wide after treatment for 24 h with auxin or dTAG-13 (Fig. 1d,e and Extended Data Fig. 1e–h). Consistently, H3K4me3 exhibited a global decrease on average across all TSS regions genome-wide within 2 h of auxin/dTAG-13 treatment. These effects on H3K4 methylation were further validated by analysis using ChIP with quantitative PCR (ChIP–qPCR) (Extended Data Fig. 1i–l). The loss of DPY30 and RBBP5 prevented long-term subcloning of the degron cells, whereas their acute degradation led to slower proliferation (Extended Data Fig. 2a–c).

Rapid turnover of H3K4me3 by KDM5

The rapid loss of H3K4me3 suggests that H3K4me3 levels are dynamically regulated by histone turnover and/or active demethylation. To investigate whether the rapid turnover of H3K4me3 was due to the activities of the KDM5 H3K4me3/me2 demethylases[22], we deleted *Kdm5a* and *Kdm5b* (*Kdm5a/b* double knockout (dKO)) in the DPY30–mAID mES cell line (DPY30–mAID *Kdm5a/b* dKO) (Fig. 1f). The deletion of the two genes did not lead to significant changes in cell proliferation and expression of pluripotent genes in the two independent isolated clones (Extended Data Fig. 2d,e). The dKO cells led to a global increase in H3K4me3 levels at steady state (Fig. 1g), consistent with previously reported data[22,23]. In contrast to in DPY30–mAID cells, in which global H3K4me3 was lost within 2 h, global H3K4me3 persisted for 8 h in dKO cells. By contrast, the turnover patterns of H3K4me1 and H3K4me2 were largely unaffected in the dKO cells (Fig. 1h). Furthermore, the *Kdm5a/b*-dKO cells showed a delayed reduction in H3K4me3 genome-wide relative to DPY30–mAID cells (Fig. 1e,i). Moreover, we found that the binding of the catalytic and non-catalytic components of the SET/COMPASS complexes was not significantly decreased after degradation of DPY30 (Extended Data Fig. 2f). These data demonstrate that the rapid turnover of H3K4me3 is dependent on KDM5A and KDM5B in mES cells.

H3K4me3 is required for transcription

To determine the primary effects on transcription after rapid H3K4me3 removal, we measured the synthesis of newly transcribed RNAs using thiol(SH)-linked alkylation for the metabolic sequencing of RNA (SLAM-seq)[24] in the degron cell lines (Extended Data Fig. 3a–d). Short-term treatment with auxin and dTAG-13 (2 h and 8 h) led to a significant reduction in mRNA synthesis (Fig. 2a,b). At these time points, H3K4me3 was lost, whereas there was a minimal effect on H3K4me2 and H3K4me1 levels (Fig. 1b–e). The number of downregulated genes ($P < 0.05$ and $\log_2$-transformed fold change < -1), as well as their magnitude of decreased expression, increased over time in both degron systems (Fig. 2a,b). Notably, the earliest downregulated genes were more likely to be regulated by a CGI-rich promoter, to show higher levels and broader peaks of H3K4me3 than genes with unaltered expression, and to be involved in processes such as ribosome biogenesis, translation and cell cycle progression (Extended Data Fig. 3e–h). Taken together, as only H3K4me3 (and not H3K4me2/me1) was affected at early timepoints after acute loss of DPY30 or RBBP5, these data suggest that H3K4me3 is required for transcription.

If H3K4me3 were the determining factor regulating transcription, deletion of *Kdm5a* and *Kdm5b* should also lead to a significant delay in gene expression changes in response to DPY30 loss. To test this, we performed SLAM-seq analysis of auxin-treated dKO cells (Fig. 2c). Indeed, we observed a significant delay in the decrease in mRNA synthesis in the dKO cells compared with in the DPY30–mAID cells (Fig. 2d,e). Specifically, only 41 and 186 genes were downregulated in the dKO cells 2 h and 8 h after auxin treatment, respectively, whereas 379 and 1,115 downregulated genes were found at the same timepoints in DPY30–mAID cells after auxin treatment, respectively (Fig. 2f). Thus, these results suggest that the reduction in H3K4me3 levels contributes to the observed decrease in transcription.

Intact PIC formation after loss of H3K4me3

As H3K4me3 has been reported to recruit proteins to enhance transcription initiation[1,25], we examined whether the loss of H3K4me3 could lead to changes in the expression of proteins in the RNAPII pre-initiation complex (PIC). As demonstrated in Fig. 3a, acute loss of H3K4me3 did not affect the global protein levels of multiple subunits of the PIC, such as CDK7 (TFIIH) and TBP (TFIID). Furthermore, reported H3K4me3-binding proteins also showed comparable protein levels and genome-wide binding patterns in control and auxin/dTAG-13-treated cells (Fig. 3a and Extended Data Fig. 4a). To address whether H3K4me3 loss would lead to changes in RNAPII PIC formation, we affinity-purified proteins that associate with RNAPII and determined their identities using mass spectrometry (MS) analysis of samples in which DPY30 was depleted or not (Extended Data Fig. 4b,c). However, loss of DPY30 did not lead to significant changes in the expression of RNAPII-associated proteins or in their interaction with RNAPII (Extended Data Fig. 4d and Supplementary Table 3). Thus, the loss of H3K4me3 does not lead to detectable changes in the formation of the RNAPII PIC or the association with PIC subunits to TSSs.

H3K4me3 regulates RNAPII occupancy

To determine the relationship between H3K4me3 and RNAPII occupancy genome-wide, we performed a ChIP–seq analysis of total RNAPII in the degron cells. Consistent with the lack of detectable effects on RNAPII PIC formation, we did not observe decreased RNAPII enrichments at promoter regions after auxin or dTAG-13 treatment. By contrast, we found an increase in RNAPII occupancy at promoter regions (Extended Data Fig. 4e), which indicated that acute loss of H3K4me3 promoted RNAPII pausing. This was confirmed by calculating the RNAPII pausing index[26] in both DPY30–mAID and RBBP5–FKBP cells (Fig. 3b,c and Extended Data Fig. 4f,g). Consistent with the increased pausing, we observed an accumulation of the RNAPII CTD Ser5 phosphorylation (Ser 5p) and negative elongation factor A (NELFA) at promoter-proximal regions and a marked reduction in RNAPII CTD Ser2 phosphorylation (Ser 2p) throughout gene bodies (Extended Data Fig. 4h,i). These data suggest that H3K4me3 is involved in the regulation of RNAPII pausing.

To gain further support that the observed effect on RNAPII pause-release was caused by loss of H3K4me3, we determined the effect of degrading DPY30 in *Kdm5a/b*-dKO cells. We confirmed that KDM5A and KDM5B were lost from TSSs in the dKO cells using ChIP–seq, and that the loss of the KDM5s led to a corresponding increase in H3K4me3 levels under steady-state conditions (Fig. 3d). Moreover, we found that the increase in H3K4me3 corresponds to a concomitant decrease in steady-state RNAPII pausing in dKO cells (Fig. 3e and Extended Data Fig. 5a). We also observed a significant delay in the onset of RNAPII pausing in the dKO cells compared with in DPY30–mAID cells expressing KDM5A/B (Fig. 3e and Extended Data Fig. 5a–c). Moreover, we performed nascent transcription with mammalian native elongating transcript sequencing (mNET–seq)[27] at single-nucleotide resolution in the DPY30–mAID and dKO lines. Consistently, the mNET–seq data showed that loss of KDM5A/B led to a significant delay in RNAPII pausing compared with in KDM5A/B-expressing cells (Fig. 3f). Taken together, these results show a role for H3K4me3 in regulating RNAPII pause-release.

H3K4me3 regulates RNAPII half-life

An increase in the occupancy of promoter-proximal RNAPII can be attributed to increased RNAPII initiation, blockage of RNAPII entering the gene body or both. To determine the stability and half-life of RNAPII at promoter-proximal regions, we inhibited transcription initiation with triptolide[28–30] and combined it with mNET–seq (Fig. 3g). The single-nucleotide resolution of the mNET–seq data enabled us to specifically measure the dynamics of RNAPII-engaged RNA at

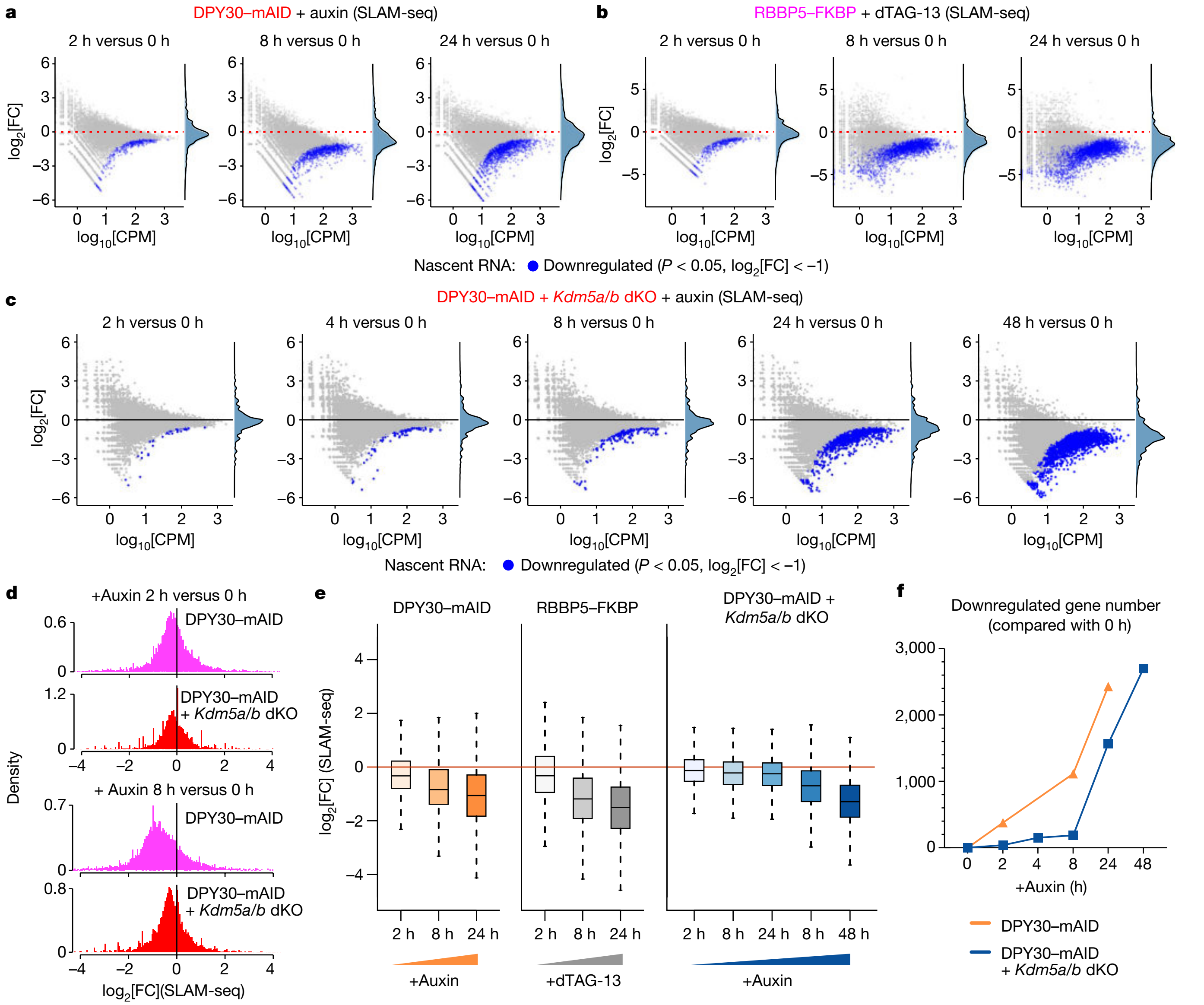


Fig. 2 | H3K4me3 is required for nascent transcription. a, *MA* plots depicting changes in nascent transcription (SLAM-seq) at the indicated times after auxin treatment in DPY30–mAID cells. $n = 3$ biological replicates. CPM, counts per million mapped reads; FC, fold change. Adjusted *P* values were calculated using Wald tests in DESeq2. **b**, *MA* plots depicting changes in nascent transcription (SLAM-seq) at the indicated times after dTAG-13 treatment in RBBP5–FKBP cells. $n = 3$ biological replicates. Adjusted *P* values were calculated using Wald tests in DESeq2. **c**, *MA* plots depicting changes in nascent transcription (SLAM-seq) at the indicated times after auxin treatment in DPY30–mAID *Kdm5a/b*-dKO cells. $n = 2$ biological replicates. Adjusted *P* values were calculated using Wald tests in DESeq2. **d**, The $\log_2$-transformed fold change in nascent gene expression in the depicted cell lines on the basis of the data shown in **a** and **c**. **e**, The nascent transcriptional changes ($\log_2$-transformed) for genes in indicated samples across timepoints with DPY30 or RBBP5 degradation kinetics. The box plots indicate the median (centre line), the third and first quartiles (box limits) and 1.5 × interquartile range (IQR) above and below the box (whiskers). $n = 3$ (DPY30–mAID or RBBP5–FKBP degron cells) and $n = 2$ (dKO cells) biological replicates. **f**, Comparison of the number of downregulated genes after auxin treatment for the indicated cell lines, based on the data presented in **a** and **c**.

promoter-proximal regions. We fitted the RNAPII time-course measurements to an exponential decay model[28] and calculated the half-life of promoter-paused RNAPII on all protein-coding genes (Extended Data Fig. 6a). This analysis showed that the RNAPII half-life is 5.3 min at protein-coding genes in steady-state mES cells ($n = 4{,}007$ genes) and increased to 8.96 min ($P < 2.2 \times 10^{-16}$) in auxin-treated cells (Fig. 3h and Extended Data Fig. 6b,c). This change in RNAPII half-life is similar to what has been observed after CDK9 inhibition[31]. Indeed, we found that CDK9, BRD4 and HEXIM1 occupancies were increased at the promoter regions (Extended Data Fig. 4i) in both degron cell lines. Thus, our data suggest that H3K4me3 regulates transcription by facilitating the release of paused RNAPII into productive elongation in mES cells.

H3K4me3 regulates elongation

To examine and monitor the effects of H3K4me3 on RNA synthesis rates genome-wide from actively transcribing RNAPII, we used a modified transient transcriptome sequencing (TT_{chem}-seq) method[32] with spike-in controls. Loss of H3K4me3 in auxin-treated or dTAG-13-treated cells led to a reduction of approximately 50% in the transcription of protein-coding genes at the 8 h timepoint (Fig. 4a,b) but resulted in increased RNAPII-engaged RNA at the promoter regions as measured by mNET–seq (Extended Data Fig. 6d). Furthermore, we found that there were no significant changes in RNAPII occupancy at either active or inactive enhancer regions after H3K4me3 loss (Extended Data Fig. 6e,f).

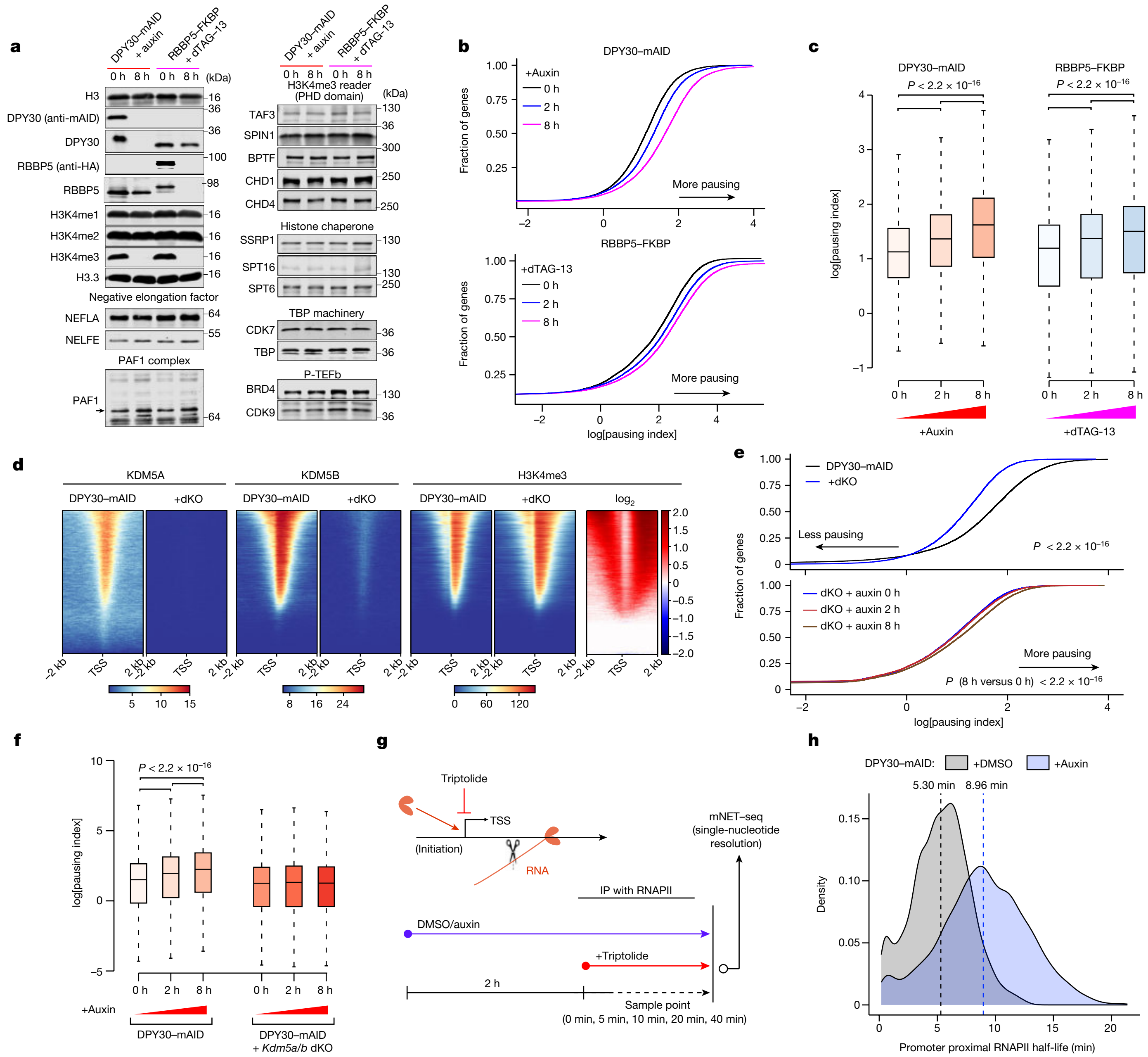


Fig. 3 | Acute loss of H3K4me3 increases the residence time of paused RNAPII. a, Immunoblot analysis of the indicated transcriptional core proteins and H3K4me3 readers in the indicated cell lines treated with or without auxin or dTAG-13 as shown. **b**, The RNAPII pausing index in control (0 h, black) and auxin-treated or dTAG-13-treated degron cells. Higher index values indicate a higher degree of RNAPII pausing. Cumulative index plots of the pausing index were calculated from total RNAPII ChIP–seq signals. **c**, The RNAPII pausing index was determined using ChIP–seq in the indicated samples with DPY30 or RBBP5 degradation kinetics. The box plots indicate the median (centre line), the third and first quartiles (box limits) and 1.5 × IQR above and below the box (whiskers). *P* values were calculated using two-sided Wilcoxon rank-sum tests. n = 12,621 genes. **d**, Comparison of the occupancy of KDM5A, KDM5B and H3K4me3 around the TSS region (TSS ± 2 kb) in DPY30–mAID and DPY30–mAID *Kdm5a/b*-dKO cells. **e**, The RNAPII pausing index in DPY30–mAID (black), DPY30–mAID *Kdm5a/b*-dKO (blue) and auxin-treated cells. Higher index values indicate a higher degree of RNAPII pausing on promoter region of genes. *P* values were calculated using two-sided Wilcoxon tests. **f**, The RNAPII pausing index determined using mNET–seq in DPY30–mAID, DPY30–mAID *Kdm5a/b*-dKO and auxin-treated cells. The box plots indicate the median (centre line), the third and first quartiles (box limits) and 1.5 × IQR above and below the box (whiskers). *P* values were calculated using two-sided Wilcoxon tests. n = 10,332 genes. **g**, The experimental strategy of the mNET–seq approach to measure the promoter-proximal RNAPII half-life after treatment with triptolide. **h**, Density plot showing increased paused RNAPII half-life of n = 4,007 genes after acute loss of H3K4me3. The average of paused RNAPII half-life is shown as a dashed line. n = 2 biological replicates.

To estimate the change in elongation velocity, we used the ratio of nascent RNA synthesis measured using TT_{chem}-seq and RNAPII RNA occupancy measured using mNET–seq (TT_{chem}-seq/mNET–seq) as a proxy for elongation velocity, which is a measurement of the amount of ongoing RNA synthesis per RNAPII molecule[33,34]. This analysis showed that acute loss of H3K4me3 caused a general transcriptome-wide decrease in elongation velocity (Fig. 4c). Consistent with this observation, the well-known epigenetic mark for transcription elongation H3K36me3 also showed a global decrease in the auxin-treated cells (Fig. 4d and Extended Data Fig. 6g).

As an orthogonal assay to investigate how H3K4me3 loss affects transcription elongation, we determined productive RNAPII elongation

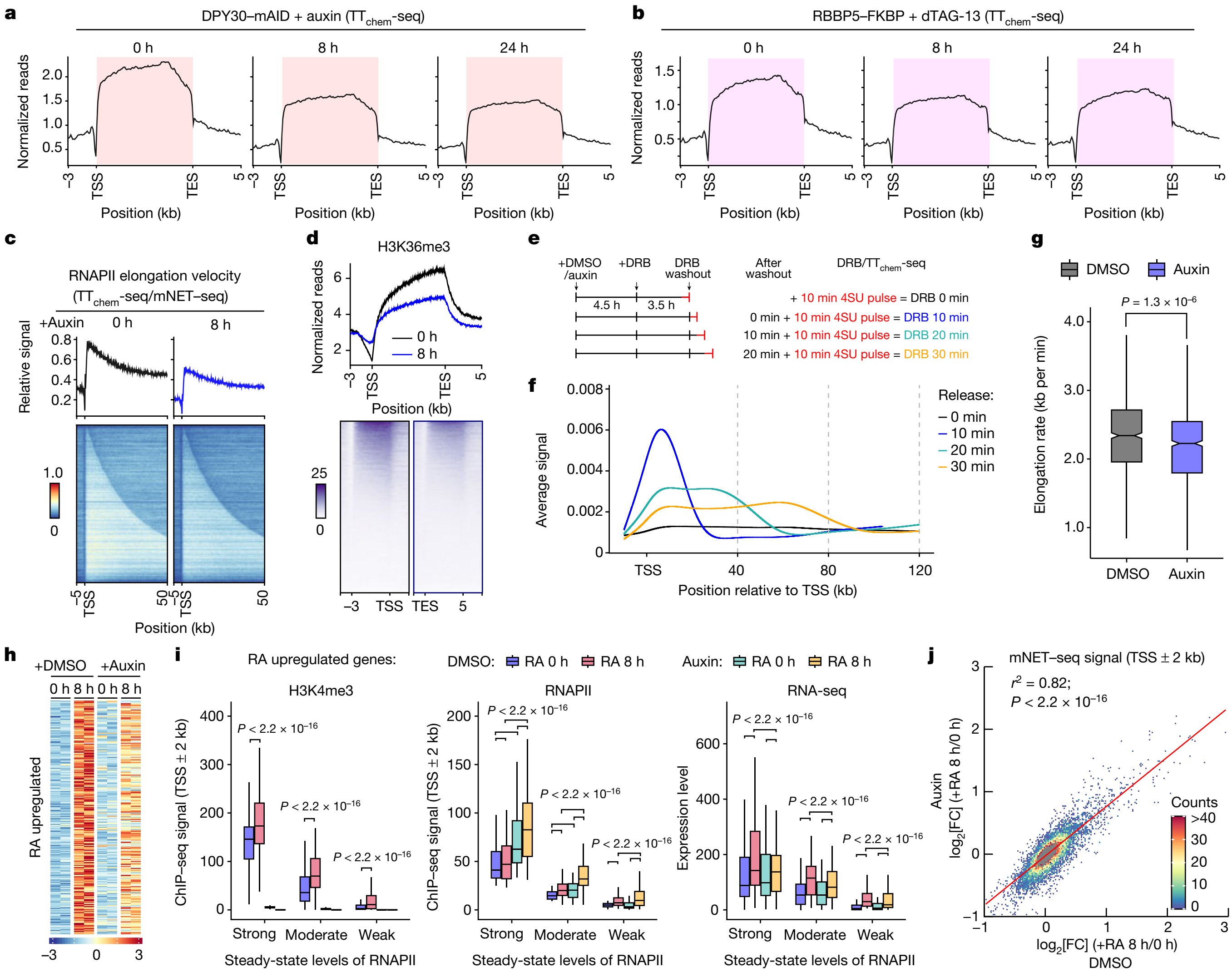


Fig. 4 | H3K4me3 regulates transcriptional elongation. a,**b**, Metagene profiles for transient transcriptome sequencing (TT$_{chem}$-seq) in control and auxin-treated (**a**) or dTAG-13-treated (**b**) cells in the indicated cell lines. TES, transcription end site. **c**, Heat maps and profiles showing changes in elongation velocities (TT$_{chem}$-seq/mNET–seq) after acute loss of H3K4me3. **d**, H3K36me3 ChIP–seq profiles and heat maps in control and auxin-treated DPY30–mAID cells. **e**, Outline of the DRB/TT$_{chem}$-seq experiment to measure RNAPII elongation rates. 4SU, 4-thiouridine. DRB 0 min, no release of DRB. **f**, DRB/TT$_{chem}$-seq metagene profiles of protein-coding genes (60–300 kb length) with non-overlapping transcriptional units (n = 3,566) in the depicted cells. Lines are computationally fitted splines. **g**, Box plot showing decreased RNAPII elongation rates after H3K4me3 loss. P values were calculated using two-sided Wilcoxon tests. n = 855 genes with RPM > 100. The box plots indicate the median (centre line), the third and first quartiles (box limits) and 1.5 × IQR above and below the box (whiskers). **h**, The upregulated genes response to RA treatment in auxin-treated and DMSO-treated cells (n = 2). Gene expression is shown as relative Z-scores across the samples. **i**, The changes in H3K4me3 and RNAPII ChIP–seq at TSSs (±2 kb), and RNA-sequencing analysis of RA-response genes (upregulated genes) in the indicated samples. The box plots indicate the median (centre line), the third and first quartiles (box limits) and 1.5 × IQR above and below the box (whiskers). P values were calculated using two-sided Wilcoxon tests. n = 77 genes for each group. **j**, The correlation of mNET–seq signal around the TSS region (TSS ± 2 kb) at 8 h after RA treatment with or without H3K4me3. Pearson correlation and P values are reported at the top. P values were calculated using two-sided Wilcoxon tests.

rates using TT$_{chem}$-seq in combination with the reversible CDK9 inhibitor 5,6-dichlorobenzimidazole 1-β-d-ribofuranoside (DRB) (DRB/TT$_{chem}$-seq)[32]. The progression of RNAPII was reflected by metagene coverage profiles after DRB release and clear transcription wave peaks on genes (Fig. 4e,f). The average elongation rates for the 3,655 non-overlapping protein-coding genes (60 kb to 300 kb in length) were 2.2 kb per min in control cells (Extended Data Fig. 6h), which is in good agreement with data obtained in HEK293[32] and HCT116[35] cells. We found that loss of H3K4me3 led to a significant decrease in RNAPII elongation rates (Fig. 4g and Extended Data Fig. 6i,j). Taken together, these observations suggest that H3K4me3 regulates both RNAPII pause-release and elongation in mES cells.

Role of H3K4me3 in initiation

It has previously been suggested, mainly on the basis of in vitro experiments[1,25], that H3K4me3 functions to facilitate the formation of the PIC at TSSs. To test this hypothesis, we determined whether H3K4me3 is required for de novo activation of transcription in response to retinoic acid (RA)-induced differentiation (Extended Data Fig. 7a). A principal component analysis and correlation analysis of the entire transcriptome showed that loss of H3K4me3 did not impair the overall rewiring of gene expression induced by RA treatment (Fig. 4h and Extended Data Fig. 7b–e), suggesting that H3K4me3 is dispensable for the transcription initiation of these genes.

Considering that many RA-response genes are bivalent with both low levels of gene expression and the histone modifications H3K4me3 and H3K27me3[36] at steady state, they may already be primed for de novo transcription. To analyse this in more detail, we determined H3K4me3 and RNAPII location using ChIP–seq analysis of DPY30–mAID cells after RA treatment with or without auxin treatment. As expected, H3K4me3 was lost in auxin-treated cells and showed increased RNAPII at promoter regions (Extended Data Fig. 6f,g). By dividing the RA-induced genes into three categories on the basis of the levels of RNAPII enrichment (strong, moderate and weak) at promoter regions in steady-state mES cells (Fig. 4i), we found increased RNAPII enrichments and gene expression for the same three category genes in the auxin-treated cells (Fig. 4i), suggesting that the RA-response genes can be initiated de novo in the absence of H3K4me3. As these steady-state data cannot rule out an effect on transcriptional initiation, high-resolution mNET–seq was performed. Correlation analysis of the mNET–seq results also showed that the loss of H3K4me3 did not impair the loading of RNAPII at the promoter-proximal region of genes induced by RA treatment (Fig. 4j). Thus, we conclude that H3K4me3 is not required for RNAPII loading and for transcriptional initiation.

The H3K4me3-dependent RNAPII interactome

To understand the mechanism leading to increased promoter-proximal pausing in response to H3K4me3 loss, we combined CRISPR-based genome editing, APEX2-based proximity labelling[37] and quantitative MS (SILAC/MS) to obtain a high-resolution view of the molecular and spatial organization of RNAPII with or without H3K4me3 (Fig. 5a). We knocked-in APEX2–Flag-tagged *Rbp1* in both degron cell lines, which did not lead to detectable effects on the cellular functions of the protein (Extended Data Fig. 8a–e). After activation with H_2O_2, APEX2 oxidizes phenol derivatives (biotin-phenol), which covalently react with the nearby endogenous proteins (Extended Data Fig. 8f,g). To capture the changes on chromatin after the loss of H3K4me3, we isolated the chromatin fraction from each sample before mixing the light and heavy conditions for MS (Extended Data Fig. 8h). We identified 1,901 proteins that were in the proximity of RPB1, and KEGG and Gene Ontology analyses showed that most of these proteins are associated with RNA biogenesis processes (Extended Data Fig. 8i–l and Supplementary Table 4), highlighting the quality of the RPB1–APEX2 data.

We focused on the changes in the two early timepoints (2 h and 8 h) for which only H3K4me3 (and not H3K4me2 and H3K4me1) was lost in response to auxin treatment (Extended Data Fig. 8m). Importantly, a large proportion of the proteins that showed differential interactions with RPB1 after DPY30 degradation were shared at the two times (Extended Data Fig. 8n). Consistent with the experimental design, DPY30 was the top downregulated protein in all of the auxin-treated samples (Extended Data Fig. 8o,p). Moreover, functional enrichment analysis of the common 228 downregulated proteins indicated a strong enrichment for mRNA-processing components of the core RNAPII machinery (Extended Data Fig. 8q). In addition to H3K4me3 loss in the cells, terms such as 'positive regulation of histone H3K4 methylation', 'chromatin binding' and 'histone binding' were more prominently decreased with chromatin RNAPII (Extended Data Fig. 8q).

H3K4me3-dependent recruitment of INTS11

To identify proteins that could potentially explain the requirement for H3K4me3 in promoter-proximal pause-release, we overlapped the common downregulated proteins with proteins that have previously been shown to associate with H3K4me3 using cross-linked ChIP–MS[38]. Notably, three proteins were in common between the two protein groups: DPY30 itself, PAF1 and INTS11 (Extended Data Fig. 9a). These three proteins were also found to be preferentially enriched in the H3K4me3 ChIP–MS data, when compared with ChIP–MS data from other heterochromatin or non-promoter specific histone modifications (Extended Data Fig. 9b). PAF1[39,40] and INTS11[41,42] have both been reported to have important roles in RNAPII pause-release and also in transcriptional elongation. We chose to focus on INTS11—an endonuclease subunit of the Integrator complex—as it showed a more substantial decrease in RPB1 association than PAF1 in response to H3K4me3 depletion (Extended Data Fig. 8p).

The reduction in RNAPII-bound INTS11 after acute loss of H3K4me3 was further confirmed by western blot analysis of both DPY30–mAID and RBBP5–FKBP degron cells (Fig. 5b and Extended Data Fig. 9c). Recent studies indicate that integrator primarily attenuates gene expression in *Drosophila*[43] and in human HeLa[44] cells, while other studies have shown a critical role of INTS11 in gene activation and elongation[45,46]. To improve our understanding of the role of INTS11 in transcription regulation, we generated INTS11–FKBP-knockin cells along with a double haemagglutinin (HA) epitope tag in the DPY30–mAID RPB1–APEX2 degron mouse ES cell line (Extended Data Fig. 9d). Western blot analysis detected similar levels of INTS11 in the constructed cell line and in non-tagged cells, and the tagging of INTS11 did not lead to detectable effects on the expression of pluripotency genes and cell proliferation (Extended Data Fig. 9e–g). The addition of dTAG-13 to these cells led to rapid degradation of the INTS11–FKBP–HA fusion protein within 2 h (Fig. 5c). Consistent with our findings for the role of H3K4me3 in regulating RNAPII pause-release, acute loss of INTS11 led to increased RNAPII pausing at promoter-proximal regions of genes (Fig. 5d–f) and to inhibition of cell proliferation (Extended Data Fig. 9g). Moreover, the enrichment of INTS11 at TSSs was significantly reduced in response to DPY30 degradation (Extended Data Fig. 9h). These data suggest H3K4me3-dependent recruitment of INTS11, which is required to enable transcriptional pause-release.

We further studied the effects of INTS11 loss by performing TT_{chem}-seq analysis of INTS11 degron cells (Extended Data Fig. 9i). Loss of INTS11 led to a significant increase in the RNA synthesis and RNAPII of non-coding short transcripts, such as upstream antisense RNAs (uaRNAs/PROMPTs) (Extended Data Fig. 9j,k). By contrast, a global decrease in the expression of protein-coding genes (Fig. 5g), especially at the pausing sites between TSS and the first (+1) nucleosome (Fig. 5h) was observed after INTS11 degradation. Analysis of elongation velocity at protein-coding genes showed broadly decreased productive elongation after INTS11 loss (Fig. 5i). Notably, the degradation of INTS11 led to the loss of only part of the integrator complex from chromatin (Fig. 5j). Taken together, these findings further support a functional link between H3K4me3 and INTS11.

To further investigate this link, we measured ongoing transcription using SLAM-seq analysis of the INTS11–FKBP degron cell line (Extended Data Fig. 10a). Notably, short-term treatment with dTAG-13 led to a significant decrease in nascent mRNA synthesis (Fig. 5k and Extended Data Fig. 10b). We also observed a strong correlation between the enrichments of H3K4me3 and INTS11 at TSS promoter-proximal regions (Fig. 5l) and that there was a significant enrichment of H3K4me3 at INTS11-bound genes, but not at INTS11-unbound genes (Fig. 5m and Extended Data Fig. 10c,d). By plotting the fold changes in nascent RNA expression determined in DPY30–mAID cells in response to auxin treatment and in INTS11–FKBP cells treated with dTAG-13, we also showed that most of the changes were correlated between the two degron systems (Fig. 5n).

Finally, we investigated whether similar correlations exist in other cell types and analysed published data from different human cell lines—HeLa[42], THP1[47] and HL60[48] (Supplementary Fig. 1a). ChIP–seq experiments in HeLa, THP1 and HL60 cells identified that ~60%, ~52% and ~61% of INTS11 peaks overlapped with promoters, respectively (Supplementary Fig. 1b). Notably, compared with the H3K4me1-occupied regions, most of the H3K4me3-occupied regions were marked with significant INTS11 and RNAPII levels (Supplementary Fig. 2a). In support of the results obtained in mES cells, INTS11 occupancy was heavily enriched towards the promoter-proximal region of genes in all of

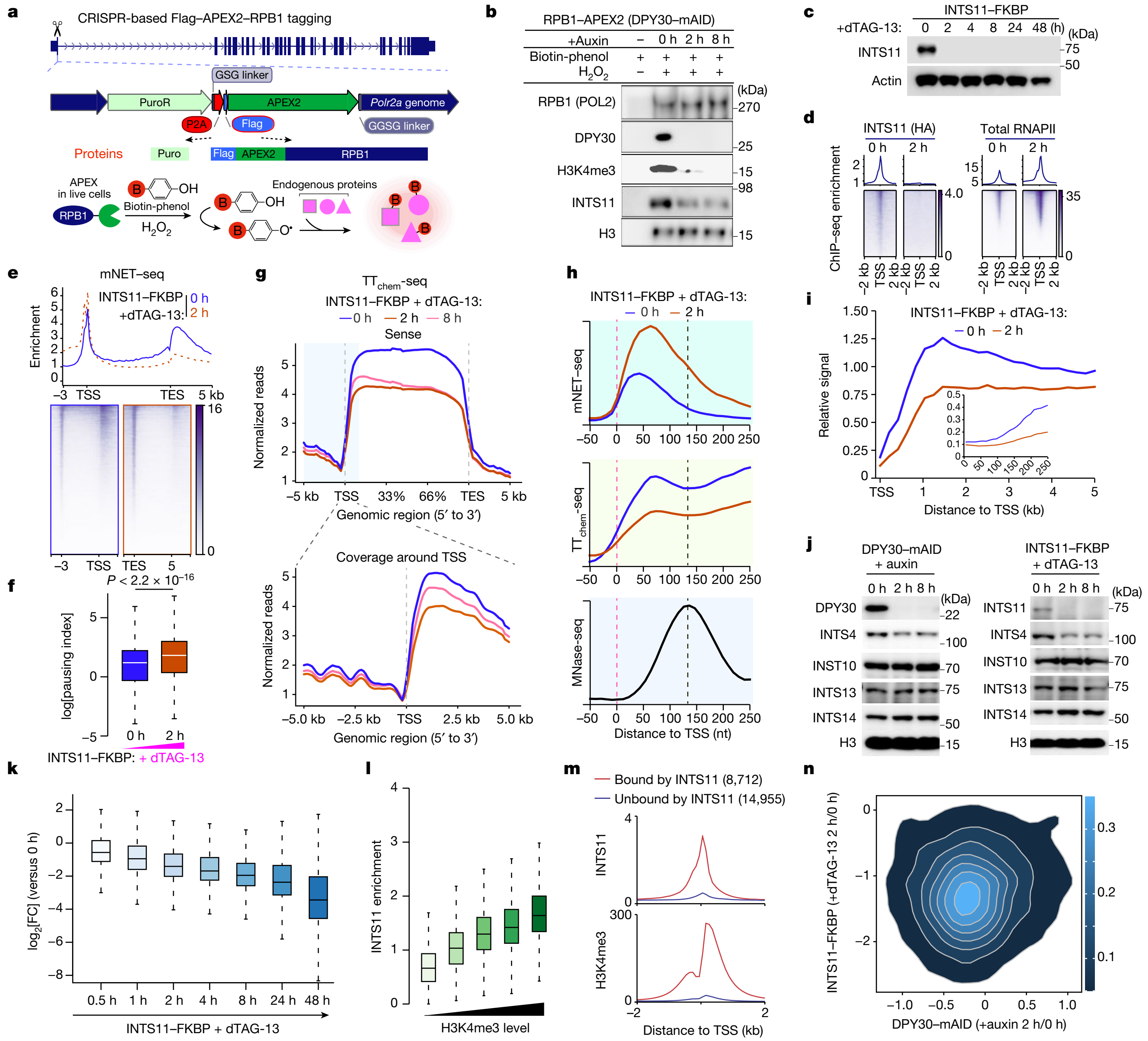


Fig. 5 | INTS11 regulates pause-release and transcription dependent on H3K4me3. a, The strategy for CRISPR-based Flag–APEX2–RPB1 (RNAPII–APEX2) tagging. **b**, Validation of H3K4me3-dependent INTS11 chromatin interaction in DPY30–mAID cells. **c**, Western blot analysis of HA-tagged INTS11 and actin in INTS11–FKBP cells. **d**, HA-tagged INTS11 and total RNAPII ChIP–seq profiles and heat maps in INTS11–FKBP cells. **e**, mNET–seq profiles and heat maps in INTS11–FKBP degron cells. **f**, The RNAPII pausing index was determined using mNET–seq in INTS11–FKBP cells. The box plots indicate the median (centre line), the third and first quartiles (box limits) and 1.5 × IQR above and below the box (whiskers). *P* values were calculated using two-sided Wilcoxon tests. *n* = 10,332 genes. **g**, Metagene transcriptional profiles were acquired using TT_{chem}-seq in INTS11–FKBP degron cells. **h**, Metagene analyses of mNET–seq and TT_{chem}-seq signals at single-nucleotide resolution acquired in INTS11–FKBP cells. MNase-seq, micrococcal nuclease sequencing. **i**, Analysis of elongation velocities (TT_{chem}-seq/mNET–seq) after acute loss of INTS11. **j**, Immunoblot analysis of integrator subunits in the indicated cell lines treated with or without auxin or dTAG-13 as shown. **k**, Box plot comparing the log_2-transformed fold change in SLAM-seq at the indicated timepoints. The box plots indicate the median (centre line), the third and first quartiles (box limits) and 1.5 × IQR above and below the box (whiskers). *n* = 13,776 genes. **l**, The INTS11 signal at TSSs (± 2 kb) in the indicated groups. The box plots indicate the median (centre line), the third and first quartiles (box limits) and 1.5 × IQR above and below the box (whiskers). *n* = 4,700 genes. **m**, The average distribution of INTS11 and H3K4me3 ChIP–seq signals at INTS11-bound genes (*n* = 8,712) versus INTS11-unbound genes (*n* = 14,955) in mES cells. **n**, 2D kernel density plot showing the relationship between SLAM-seq changes in INTS11–FKBP and DPY30–mAID degron cells. The colour bar reflects the intensity.

these human cell lines (Supplementary Fig. 2b). Moreover, heat maps also demonstrated a significant colocalization of the peak binding sites for INTS11 and RNAPII with the peak sites of H3K4me3 enrichment on active promoters (Supplementary Fig. 2b). Thus, these published ChIP–seq results from different cells validated a correlation between INTS11-binding sites and sites of H3K4me3 enrichment across the genome. Taken together, these results show that H3K4me3 regulates promotor-proximal pausing through a mechanism involving the recruitment of the INTS11, which is essential for the eviction of paused RNAPII and transcriptional elongation.

Discussion

H3K4me3, which is catalysed by the SET1/COMPASS complexes, is tightly associated with TSSs and is widely believed to be involved in regulating transcription initiation[1,2,7,25]. By generating and using model systems to study the direct role of H3K4me3 in regulating transcription, we have shown that loss of H3K4me3 leads to an increase in RNAPII at promoters. Moreover, we have shown, through multiple orthogonal assays, that H3K4me3 regulates RNAPII pausing and potentially also has a role in elongation. A connection between H3K4me3 and elongation has previously been suggested on the basis of results in cancer cells[12] and plants[49]. Notably, we did not detect a role for H3K4me3 in transcriptional initiation and activation of de novo transcribed genes in response to RA-induced differentiation. Furthermore, we observed a fast turnover of H3K4me3 that is dependent on KDM5 demethylase activity, and we propose that this rapid turnover of H3K4me3 is important for the dynamic regulation of transcriptional output. Our experimental approach does not enable us to delineate specific roles for H3K4me1 and H3K4me2 in transcriptional regulation because all H3K4 methylation states are influenced at later time points. Notably, recent results suggest that SET1A/B can also regulate CpG-island-associated gene expression independently of H3K4me3 and its methyltransferase activity[50], indicating potential distinct roles of components of the SET1/COMPASS complexes.

Paused RNAPII has been proposed to be coupled with transcription and mRNA-processing events by helping to maintain the accessibility of promoters for regulatory factors that activate transcription by recruiting P-TEFb and other factors, and promoting the release of paused RNAPII into gene bodies[30,33,39]. Whether histone modifications have a role in regulating the RNAPII pause-release step is not well understood. Our results show that H3K4me3 coupled with INTS11 is involved in regulating RNAPII promoter-proximal pausing. We propose a model in which H3K4me3 regulates transcriptional cycles by facilitating the recruitment of INTS11 to protein-coding genes (Extended Data Fig. 11). INTS11 can subsequently, through its RNA endonuclease activity, mediate productive transcriptional elongation by evicting paused RNAPII. This model is supported by recent studies showing that mammalian INTS11 facilitates RNAPII pause-release and gene expression[45,51]. Most of our conclusions are also supported by another recent study; however, in this study the authors observed an upregulation of protein-coding short transcripts after INTS11 downregulation[52]. Although we do not know the reason for the different observations, it may reflect the efficiency of the two different degron systems used in the studies.

Further work will be needed to determine the detailed molecular mechanism that leads to the recruitment of INTS11 and its function in elongation. In summary, our study demonstrates that H3K4me3 is indeed an important post-translational modification that regulates promoter-proximal pause-release and facilitates gene expression.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-05780-8.

1. Lauberth, S. M. et al. H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. *Cell* **152**, 1021–1036 (2013).
2. Shilatifard, A. The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis. *Annu. Rev. Biochem.* **81**, 65–95 (2012).
3. Miller, T. et al. COMPASS: a complex of proteins associated with a trithorax-related SET domain protein. *Proc. Natl Acad. Sci. USA* **98**, 12902–12907 (2001).
4. Qu, Q. et al. Structure and conformational dynamics of a COMPASS histone H3K4 methyltransferase complex. *Cell* **174**, 1117–1126 (2018).
5. Piunti, A. & Shilatifard, A. Epigenetic balance of gene expression by Polycomb and COMPASS families. *Science* **352**, aad9780 (2016).
6. Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. *Science* **357**, eaal2380 (2017).
7. Rao, R. C. & Dou, Y. Hijacked in cancer: the KMT2 (MLL) family of methyltransferases. *Nat. Rev. Cancer* **15**, 334–346 (2015).
8. Vermeulen, M. et al. Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. *Cell* **142**, 967–980 (2010).
9. Ooi, S. K. et al. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* **448**, 714–717 (2007).
10. Liu, X. Y. et al. Distinct features of H3K4me3 and H3K27me3 chromatin domains in pre-implantation embryos. *Nature* **537**, 558–562 (2016).
11. Sankar, A. et al. KDM4A regulates the maternal-to-zygotic transition by protecting broad H3K4me3 domains from H3K9me3 invasion in oocytes. *Nat. Cell Biol.* **22**, 380–388 (2020).
12. Chen, K. F. et al. Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes. *Nat. Genet.* **47**, 1149–1157 (2015).
13. Clouaire, T. et al. Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Genes Dev.* **26**, 1714–1728 (2012).
14. Ang, Y. S. et al. Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. *Cell* **145**, 183–197 (2011).
15. Jiang, H. et al. Role for Dpy-30 in ES cell-fate specification by regulation of H3K4 methylation within bivalent domains. *Cell* **144**, 513–525 (2011).
16. Denissov, S. et al. Mll2 is required for H3K4 trimethylation on bivalent promoters in embryonic stem cells, whereas Mll1 is redundant. *Development* **141**, 526–537 (2014).
17. Dorighi, K. M. et al. Mll3 and Mll4 facilitate enhancer RNA synthesis and transcription from promoters independently of H3K4 monomethylation. *Mol. Cell* **66**, 568–576 (2017).
18. Bledau, A. S. et al. The H3K4 methyltransferase Setd1a is first required at the epiblast stage, whereas Setd1b becomes essential after gastrulation. *Development* **141**, 1022–1035 (2014).
19. Howe, F. S., Fischl, H., Murray, S. C. & Mellor, J. Is H3K4me3 instructive for transcription activation. *Bioessays* **39**, 1–12 (2017).
20. Natsume, T., Kiyomitsu, T., Saga, Y. & Kanemaki, M. T. Rapid protein depletion in human cells by auxin-inducible degron tagging with short homology donors. *Cell Rep.* **15**, 210–218 (2016).
21. Nabet, B. et al. The dTAG system for immediate and target-specific protein degradation. *Nat. Chem. Biol.* **14**, 431–441 (2018).
22. Christensen, J. et al. RBP2 belongs to a family of demethylases, specific for tri-and dimethylated lysine 4 on histone 3. *Cell* **128**, 1063–1076 (2007).
23. Schmitz, S. U. et al. Jarid1b targets genes regulating development and is involved in neural differentiation. *EMBO J.* **30**, 4586–4600 (2011).
24. Herzog, V. A. et al. Thiol-linked alkylation of RNA to assess expression dynamics. *Nat. Methods* **14**, 1198–1204 (2017).
25. Vermeulen, M. et al. Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell* **131**, 58–69 (2007).
26. Rahl, P. B. et al. c-Myc regulates transcriptional pause release. *Cell* **141**, 432–445 (2010).
27. Nojima, T., Gomes, T., Carmo-Fonseca, M. & Proudfoot, N. J. Mammalian NET-seq analysis defines nascent RNA profiles and associated RNA processing genome-wide. *Nat. Protoc.* **11**, 413–428 (2016).
28. Shao, W. & Zeitlinger, J. Paused RNA polymerase II inhibits new transcriptional initiation. *Nat. Genet.* **49**, 1045–1051 (2017).
29. Mylonas, C., Lee, C., Auld, A. L., Cisse, I. I. & Boyer, L. A. A dual role for H2A.Z.1 in modulating the dynamics of RNA polymerase II initiation and elongation. *Nat. Struct. Mol. Biol.* **28**, 435–442 (2021).
30. Jonkers, I., Kwak, H. & Lis, J. T. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *eLife* **3**, e02407 (2014).
31. Gressel, S. et al. CDK9-dependent RNA polymerase II pausing controls transcription initiation. *eLife* **6**, e29736 (2017).
32. Gregersen, L. H., Mitter, R. & Svejstrup, J. Q. Using TT_{chem}-seq for profiling nascent transcription and measuring transcript elongation. *Nat. Protoc.* **15**, 604–627 (2020).
33. Caizzi, L. et al. Efficient RNA polymerase II pause release requires U2 snRNP function. *Mol. Cell* **81**, 1920–1934 (2021).
34. Zumer, K. et al. Two distinct mechanisms of RNA polymerase II elongation stimulation in vivo. *Mol. Cell.* **81**, 3096–3109 (2021).
35. Geijer, M. E. et al. Elongation factor ELOF1 drives transcription-coupled repair and prevents genome instability. *Nat. Cell Biol.* **23**, 608–619 (2021).
36. Levine, M. Paused RNA polymerase II as a developmental checkpoint. *Cell* **145**, 502–511 (2011).
37. Hung, V. et al. Spatially resolved proteomic mapping in living cells with the engineered peroxidase APEX2. *Nat. Protoc.* **11**, 456–475 (2016).
38. Ji, X. et al. Chromatin proteomic profiling reveals novel proteins associated with histone-marked genomic regions. *Proc. Natl Acad. Sci. USA* **112**, 3841–3846 (2015).
39. Chen, F. X. et al. PAF1, a molecular regulator of promoter-proximal pausing by RNA polymerase II. *Cell* **162**, 1003–1015 (2015).
40. Yu, M. et al. RNA polymerase II-associated factor 1 regulates the release and phosphorylation of paused RNA polymerase II. *Science* **350**, 1383–1386 (2015).
41. Zheng, H. et al. Identification of integrator-PP2A complex (INTAC), an RNA polymerase II phosphatase. *Science* **370**, eabb5872 (2020).
42. Gardini, A. et al. Integrator regulates transcriptional initiation and pause release following activation. *Mol. Cell* **56**, 128–139 (2014).
43. Tatomer, D. C. et al. The integrator complex cleaves nascent mRNAs to attenuate transcription. *Genes Dev.* **33**, 1525–1538 (2019).
44. Elrod, N. D. et al. The integrator complex attenuates promoter-proximal transcription at protein-coding genes. *Mol. Cell* **76**, 738–752 (2019).
45. Beckedorff, F. et al. The human integrator complex facilitates transcriptional elongation by endonucleolytic cleavage of nascent transcripts. *Cell Rep.* **32**, 107917 (2020).
46. Dasilva, L. F. et al. Integrator enforces the fidelity of transcriptional termination at protein-coding genes. *Sci. Adv.* **7**, eabe3393 (2021).
47. Vervoort, S. J. et al. The PP2A–integrator–CDK9 axis fine-tunes transcription and can be targeted therapeutically in cancer. *Cell* **184**, 3143–3162 (2021).

48. Barbieri, E. et al. Targeted enhancer activation by a subunit of the integrator complex. *Mol. Cell* **71**, 103–116 (2018).
49. Copenhaver, G. P. et al. ATX1-generated H3K4me3 is required for efficient elongation of transcription, not initiation, at ATX1-regulated genes. *PLoS Genet.* **8**, e1003111 (2012).
50. Hughes, A. L. et al. A CpG island-encoded mechanism protects genes from premature transcription termination. *Nat. Commun.***14**, 726 (2023).
51. Zheng, H. et al. Structural basis of INTAC-regulated transcription. Preprint at *bioRxiv* https://doi.org/10.1101/2021.11.29.470345 (2021).
52. Stein, C. B. et al. Integrator endonuclease drives promoter-proximal termination at all RNA polymerase II-transcribed loci. *Mol. Cell* **82**, 4232–4245 (2022).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

Methods

mES cell culture and differentiation

E14 ES cells (129/Ola background) were maintained on 0.2% gelatin-coated plates in Glasgow minimum essential medium (GMEM, Sigma-Aldrich, G5154) containing 15% fetal bovine serum (Gibco, 26140079), supplemented with 1× penicillin–streptomycin (Thermo Fisher Scientific, 15140122), 2 mM GlutaMax (Thermo Fisher Scientific, 35050061), 50 μM β-mercaptoethanol (Thermo Fisher Scientific, 21985023), 0.1 mM non-essential amino acids (Thermo Fisher Scientific, 11140050), 1 mM sodium pyruvate (Thermo Fisher Scientific, 11360070) and Leukaemia inhibitory factor (LIF, 1,000 U ml^{-1}, Millipore), referred to as serum mES cell medium. Cells were passaged every 2 days by aspirating the medium, dissociating the cells with trypsin/EDTA solution (TE) briefly at room temperature before rinsing and dissociation in mES cell medium by pipetting. Cells were pelleted by centrifugation at 300*g* for 5 min. mES cell transfection was performed using Lipofectamine 3000 (Thermo Fisher Scientific, L3000-001) according to the manufacturer's instructions. Cell counts were performed using the Countess II automated cell counter (Thermo Fisher Scientific, AMQAX1000) using 10 μl of cell suspension and 10 μl of Gibco Trypan Blue Solution (Gibco, 15250061) according to manufacturer's instructions. For all-*trans*-retinoic acid (RA, Sigma-Aldrich, R2625-50MG) treatment, cells were induced with 1 μM RA and without LIF for the indicated time. During differentiation, RA medium was changed every 24 h. For SILAC experiments, cells were cultured in SILAC DMEM (Thermo Fisher Scientific, A33822) containing 15% dialysed FBS (Thermo Fisher Scientific, 26400044), to which either $^{13}C_6{}^{15}N_2$ L-lysine-2HCl (Thermo Fisher Scientific, 88209) and $^{13}C_6{}^{15}N_4$ L-arginine-HCl (Thermo Fisher Scientific, 89990) (heavy), or L-lysine (Sigma-Aldrich, L8662) and L-arginine (Sigma-Aldrich, L8094) containing only light isotopes (light) was added. All cell lines were subjected to STR authentification through ATCC and were tested for mycoplasma contamination.

Generation of mES cell knockin cell lines

For the generation of the auxin-inducible degradation system for DPY30 sgRNA targeting the stop-codon region were cloned into eSpCas9(1.1)-T2A-eGFP. Left and right homology arms, as well as the mAID-T2A-BFP middle part were ligated into a modified pUC19 vector backbone (a gift from S. Pollard) using the In-Fusion cloning kit (Takara, 638910). mES cells were co-transfected with sgRNA- and donor-vector using Lipofectamine 3000 and sorted 48 h later for GFP/BFP-double-positive cells. Homozygous clones were then transfected with pPB-hygro-OsTIR1-P2A-mCherry and pBase plasmids and selected with 100 μg ml^{-1} hygromycin B (Thermo Fisher Scientific, 10687010). For the generation of the endogenous dTAG-inducible degradation system for RBBP5, sgRNA targeting the stop codon region were co-transfected into the cells and contained the following elements: left and right homology arms, as well as FKBP12(F36V), 2× HA tags, P2A and a neomycin-resistance gene. The transfected cells were selected with 100 μg ml^{-1} Geneticin selective antibiotic (G418 Sulfate) (Thermo Fisher Scientific, 10131027), single-cell sorted to obtain clonal cell lines and screened for correct biallelic integration. All homozygous insertions and knock-ins were confirmed by Sanger sequencing and western blotting. A list of the oligos and the sequences of the sgRNAs is provided in Supplementary Table 1.

Generation of mES knockout cell lines

Cells were transfected with eSpCas9(1.1)-T2A-eGFP or eSpCas9(1.1)-T2A-mCherry vectors containing a sgRNA targeting the specific genomic locus, respectively, and cells were single-cell sorted 48 h after transfection. To generate the *Kdm5a/b*-dKO cells in the DPY30–mAID cell line, we first generated the *Kdm5a*-KO in the DPY30–mAID line (*Kdm5a*$^{-/-}$) by Cas9 (sgRNAs are listed in Supplementary Table 1), then we generated the *Kdm5b* knockout in the *Kdm5a*$^{-/-}$ line (at passage three of the *Kdm5a*$^{-/-}$ line). Subsequently, two clones with both *Kdm5a* and *Kdm5b* knockout (referred to as DPY30–mAID *Kdm5a/b*-dKO in this study) were picked up for the downstream analysis. All homozygous insertions and knockouts were confirmed by Sanger sequencing and western blotting. A list of the sgRNAs is provided in Supplementary Table 1. Further characterization of the dKO cell lines showed that they did not have detectable changes in proliferation and expressed normal levels of pluripotent genes; however, as expected, they showed an increase in H3K4me3 levels as compared with the wild-type cells.

Western blotting

Cells were lysed in RIPA buffer with Halt protease inhibitor (Thermo Fisher Scientific, 78429). Proteins that were separated by SDS–PAGE using acrylamide gels (BioRad gel system) were transferred onto nitrocellulose membranes (LI-COR, 926-31092). The membranes were blocked in 5% skimmed milk (Sigma-Aldrich) in PBS-T (0.1% Tween-20 in PBS) and incubated with the primary antibody of interest (Supplementary Table 2). As secondary antibodies, either IRDye 800CW goat anti-rabbit IgG (925-32211, LI-COR Bioscience, 1:15,000) or IRDye 800CW goat anti-mouse IgG (925-32210, LI-COR Bioscience, 1:15,000) was used. Proteins were imaged using Image Studio Lite (Odyssey CLx imager, Li-COR Biosciences). Immunoblotting source data are provided in Supplementary Fig. 4.

Flow cytometry

mES cells were dissociated with trypsin/EDTA, resuspended in culture medium, centrifuged and resuspended in PBS. For intracellular flow cytometry, 0.5 ml of cold fixation buffer (BioLegend, 420801) was added and then incubated at room temperature for 10 min. Subsequently, the cells were labelled with the unconjugated rabbit DPY30 antibodies (Bethyl Laboratories, A304-296A) and subsequently with a FITC-conjugated goat anti-rabbit IgG antibody. Flow cytometry was performed using the Beckman Coulter CytoFlex system. The gating strategy is shown for the E14 sample in Supplementary Fig. 5.

RNA extraction, cDNA synthesis and RT–qPCR analysis

Total RNA was extracted using the RNeasy Plus Mini Kit (Qiagen, 74134) according to the manufacturer's protocol. One microgram of total RNA was subjected to reverse transcription using Transcriptor Universal cDNA Master (Sigma-Aldrich, 5893151001). qPCR with reverse transcription (RT–qPCR) reactions were set up in triplicate using PowerUp SYBR Green Master Mix (Thermo Fisher Scientific, A25778) and primers (listed in Supplementary Table 1). Relative quantitation was performed to a housekeeping gene using a $\Delta\Delta C_t$ method, as indicated in the corresponding figure legends. Statistical analysis was performed using GraphPad Prism v.7 (GraphPad).

3′-RNA Quant-seq and SLAM-seq

Cells (1×10^6 per treatment condition) were resuspended in 350 μl of buffer RLT plus, and total RNA was extracted from cell pellets using the RNeasy Plus Mini kit (Qiagen, 74134). For SLAM-seq experiments[24], cells were incubated with 100 μM 4-thiouridine (4SU; Biosynth, NT06186) for 60 min before RNA isolation. RNA (1 μg) was treated with 10 mM iodoacetamide in a 50 μl reaction volume at 50 °C for 15 min with 50 mM NaH_2PO_4 (pH 8.0), and 50% (v/v) DMSO followed by addition of 1 μl of 1 M dithiothreitol (DTT) to stop the reaction. RNA was precipitated at −80 °C for 60 min with 1 μl of GlycoBlue (Thermo Fisher Scientific, AM9515), 5 μl of 3 M sodium acetate (pH 5.2) and 180 μl of ethanol (≥99.0%). RNA was pelleted at 4 °C (12,000*g* for 30 min), washed with 1 ml of 80% ethanol and centrifuged at 4 °C (12,000*g* for 30 min). The RNA pellet was dried at room temperature for 10 min and resuspended in 20 μl of nuclease-free H_2O. RNA yield and quality were assessed using the 2200 TapeStation (Agilent). Sequencing libraries were prepared using the QuantSeq 3′-mRNA Seq Library Prep Kit FWD for Illumina (Lexogen, SKU: 015.96) from 500 ng or 100 ng of total RNA

spiked with ERCC RNA Spike-In Mix 1 (1:1,000, Thermo Fisher Scientific, 4456740). In brief, first-strand (oligo(dT)) cDNA synthesis was followed by RNA removal and second-strand synthesis by random priming. The double-stranded library was bead-purified to remove reaction components before PCR amplification with i7 single-index primers for 10 cycles. Amplified libraries were again bead-purified according to the manufacturer's protocol, and the concentration was measured by Qubit assay. All of the samples were checked for fragment size distribution on the TapeStation before pooling for 75 bp or 45 bp single-end read sequencing on the Illumina NextSeq 550 platform.

ChIP–seq

ChIP experiments were performed according to a standard protocol. In brief, ES cells were cross-linked by the addition of 1% formaldehyde (Sigma-Aldrich, 252549-1L) in the dish for 10 min at room temperature before quenching with 0.125 M glycine. The fixed cells were washed with PBS and resuspended in SDS buffer (100 mM NaCl, 50 mM Tris-HCl pH 8.0, 5 mM EDTA, 0.5% SDS, 1× protease inhibitor cocktail from Roche). The resulting nuclei were precipitated, resuspended in the immunoprecipitation buffer at 1 ml per 16 million cells (SDS buffer and Triton dilution buffer (100 mM NaCl, 100 mM Tris-HCl pH 8.0, 5 mM EDTA, 5% Triton X-100) mixed at a 2:1 ratio with the addition of 1× protease inhibitor cocktail from Roche) and processed on the Bioruptor Plus Sonicator (Diagenode) to achieve an average fragment length of 200–300 bp. Chromatin concentrations were estimated using the Nanodrop according to manufacturer's protocols. The immunoprecipitation reactions were set up in 1 ml of the immunoprecipitation buffer as indicated below and incubated overnight at 4 °C. The next day, BSA-blocked Protein G Dynabeads (Thermo Fisher Scientific, 10004D) were added to the reactions and incubated for 2 h at 4 °C. The beads were then washed three times with low-salt washing buffer (150 mM NaCl, 1% Triton X-100, 0.1% SDS, 2 mM EDTA, 20 mM Tris-HCl pH 8.0) and twice with high-salt washing buffer (500 mM NaCl, 1% Triton X-100, 0.1% SDS, 2 mM EDTA, 20 mM Tris-HCl pH 8.0). The samples were then reverse cross-linked overnight at 65 °C in the elution buffer (1% SDS, 0.1 M $NaHCO_3$) and purified using the QIAQuick PCR purification kit (Qiagen, 28506). A list of the antibodies used in this study is provided in Supplementary Table 2. Libraries for ChIP–seq were prepared using the NEBNext Ultra II DNA Library prep kit (NEB, E7645L), and AmpureXP beads (Beckman, A63881) were used for size selection. Libraries were quantified using the Qubit High Sensitivity DNA kit (Agilent, Q32854) and assessed on the TapeStation. Libraries were pooled as required, denatured and loaded onto the Illumina NextSeq 550 system with high-output kits (75 cycles). A list of all of the primers used for ChIP–qPCR is provided in Supplementary Table 1. For spiked-in ChIP–seq, 5% of the cross-linked *Drosophila* chromatin (homemade) with Spike-in Antibody (Active Motif, 61686) was added before the immunoprecipitation step according to the manufacturer's instructions.

RNAPII IP followed by MS

To measure the difference between RNAPII interactome in the presence and absence of DPY30, 1×10^8 DPY30–mAID cells were incubated with auxin for 8 h to induce DPY30 degradation (8 h samples), while 1×10^8 DPY30–mAID cells treated with DMSO served as the control (0 h samples). Cells were collected and frozen on dry ice and kept at −80 °C until immunoprecipitation (IP). The cells were thawed at 37 °C for 30 s lysed in 1.6 ml of ice cold 50 mM EPPS pH 7.5, 150 mM NaCl, 1% Triton X-100 with cOmplete, EDTA-free Protease Inhibitor Cocktail (1 tablet per 20 ml of lysis buffer), 1:100 of Sigma-Aldrich phosphatase inhibitor 2 and 3 cocktails and 250 U μl^{-1} of benzonase. Lysates were incubated on ice for 5 min to allow DNA digestion, centrifuged at 20,000*g* for 5 min to remove insoluble material and filtered through the AcroPrep 1.0 µm glass filter plate at 2,000*g* for 1 min. The concentration of protein was then estimated using the bicinchoninic acid assay. For each of the samples (DMSO- and auxin-treated cells) six immunoprecipitation reactions were performed: three with the anti-RNAPII antibody (Abcam, ab817, 8WG16) and three with IgG control (Invitrogen, 02-6102). Each reaction was performed with 1 mg of lysate at 3.3 g l^{-1} lysate concentration and 5 µg of antibody bound to protein G Sepharose (Sigma-Aldrich, 17-0618-02). The incubation was performed at 4 °C with shaking at 1,100 rpm for 1 h. The beads were then transferred to the OF 1100 filter plate (Orochem Technologies) and washed five times with ice-cold 50 mM EPPS pH 7.5, 150 mM NaCl using vacuum manifold. Then, 18 µl of 10 mM EPPS pH 8.5 with 20 ng μl^{-1} trypsin and 1 ng μl^{-1} LysC was added to the beads in each well and digestion was performed for 2 h at 37 °C at 2,000 rpm. The partial digest was then collected into a 96-well PCR plate and left overnight at room temperature to complete digestion. Then, 4 µl of 22 g l^{-1} 11 plex TMTPro tags were added to each sample. The samples were then pulled and 20 µl of the combined sample was set aside, and the rest was fractionated into six fractions using the High pH Reversed-Phase Peptide Fractionation Kit, as suggested by the manufacturer. The fractions were concatenated into four fractions (the first and fifth fractions, the second and sixth and so on were mixed) and evaporated in speed vac (0.5 µl of DMSO was added to each sample to prevent complete evaporation) and resuspended in 20 µl 0.1% TFA. For data acquisition, 4.5 µl of unfractionated sample and every fraction was analysed by using the nanoAcquity 2 µm particle size, 75 mm × 500 mm easyspray column in direct injection mode. The samples were separated using the following gradient at 300 nl min^{-1} of buffer A (0.1% formic acid in water) and buffer B (0.1% formic acid in acetonitrile): 0–7% in 10 min, 7–30% in 92 min, 30–60% in 18 min, the column was then washed with 95% B for 10 min at 400 nl min^{-1}. The column was kept at 60 °C. Eluting peptides were analysed on the Orbitrap Fusion Lumos instrument using MS3 SPS with the settings recommended by the instrument manufacturer for TMT11 plex analysis with the following modifications: (1) CID NCE for MS2 was set at 32; (2) HCD NCE for MS3 was set at 45; (3) C series exclusion was disabled as TMTPro reagent was not enabled in C-series exclusion node. The cycle time was set at 3 s and the dynamic exclusion time was set at 15 s.

APEX2-based RNAPII proximity labelling and affinity enrichment of biotinylated proteins

CRISPR–Cas9 technology was used to target endogenous *Rpb1* at the 5′ end with a cassette encoding a Flag affinity-tag and APEX2 (Flag–APEX2), resulting in RPB1 fused at its N terminus to Flag–APEX2. DPY30–mAID OsTIR1 E14 cells were co-transfected with espCas9 plasmid and a donor plasmid containing the puromycin-resistance selection gene, P2A self-cleavage site and Flag–APEX2 flanked by homology arms corresponding to the respective target genes. The sequences of the guide RNA and homology arms for targeting *Rpb1* are provided in Supplementary Table 1. The APEX2-expressing cells were incubated with 4 mM biotin-phenol reagent (Iris Biotech, LS-3500.1000) for 2 h before the start of the labelling reaction. The cells were washed with PBS (with Ca^{++} and Mg^{++}) and the labelling reaction was initiated by adding 1 mM H_2O_2 in PBS for 2 min at room temperature. The reaction was terminated by washing the cells three times with a quencher solution containing 10 mM sodium azide, 10 mM sodium ascorbate and 5 mM Trolox in PBS.

Cell fractionation for SILAC chromatin MS

Chromatin fractions were prepared as described previously[53] with some modifications. In brief, cells were lysed by swelling and mechanical force in buffer A (10 mM ammonium bicarbonate pH 8.0, 1.5 mM $MgCl_2$, 10 mM KCl, 10 mM sodium ascorbate, 5 mM Trolox, 10 mM sodium azide, 1× protease inhibitor cocktail and 0.2% NP40). Nuclei were then collected by centrifugation and chemically lysed in buffer C (20 mM ammonium bicarbonate pH 8.0, 420 mM $NaCl_2$, 20% (v/v) glycerol, 2 mM $MgCl_2$, 0.2 mM EDTA, 0.1% NP40, 10 mM sodium ascorbate, 5 mM Trolox, 10 mM sodium azide, 1× protease inhibitor cocktail and 0.5 mM DTT). Lysates were centrifuged at 20,800*g* for 45 min at 4 °C. The pellet contains the insoluble chromatin fraction and consists of DNA and

proteins tightly bound to chromatin. To solubilize the chromatin pellet, 750 U Benzonase (Sigma-Aldrich) was added, followed by 10 min incubation on ice and 5 min of agitation at room temperature. Clarified lysate was collected and the protein concentration was quantified using Bio-Rad Bradford's reagent. Approximately 4 mg lysates from SILAC heavy or light cells were mixed 1:1 and incubated with 50 µl Streptavidin magnetic beads (Pierce, 88817) at 4 °C on a rotating wheel overnight. The beads were washed four times with RIPA buffer.

Sample preparation for SILAC/MS

Eluates from biotin pull-down were transferred to fresh microfuge tubes. NuPAGE sample loading buffer was added to the beads and heated at 90 °C for 5 min. A magnetic rack was used to separate the beads from the proteins. The supernatant was then run on an SDS–PAGE gel (Bis-Tris, 4–12%) enough to get the sample into the gel. Gel sections were excised, washed, reduced with DTT, alkylated with iodoacetamide and digested overnight with trypsin at 37 °C (ref. [54]). Homemade C18 StageTips were prepared as described previously[55] and preconditioned with a 50 µl wash of methanol, 50 µl wash of 70% acetonitrile/0.1% trifluoroacetic acid and two 50 µl washes of 0.1% trifluoroacetic acid at 1,000*g*. Peptides were then loaded onto StageTips and washed with 50 µl of 0.1% formic acid and were eluted with 60 µl of 70% acetonitrile/0.1% formic acid. The samples were then vacuum centrifuged using the SpeedVac and reconstituted in 0.1% formic acid for LC–MS/MS and were analysed by microcapillary LC–MS/MS using the nanoAcquity system (Waters) with a 100 µm inner-diameter × 10 cm length C18 column (1.7 µm BEH130, Waters) configured with a 180 µm × 2 cm trap column coupled to a Q-Exactive Plus mass spectrometer (Thermo Fisher Scientific). Peptides were eluted at 300 nl min^{-1} using a 4 h acetonitrile gradient (0.1% formic acid). The Q-Exactive Plus mass spectrometer was operated in automatic, data-dependent MS/MS acquisition mode with one MS full scan (380–1,600 *m*/*z*) at 70,000 mass resolution and up to ten concurrent MS/MS scans for the ten most intense peaks selected from each survey scan. Survey scans were acquired in profile mode and MS/MS scans were acquired in centroid mode at 17,500 resolutions with an isolation window of 1.5 amu and normalized collision energy of 27; AGC was set to 1×10^6 for MS1 and 5×10^4 and 50 ms max IT for MS2; charge exclusion of unassigned, +1 and greater than 6 was enabled with dynamic exclusion of 15 s.

TT_{chem}-seq and DRB/TT_{chem}-seq analysis

We performed TT_{chem}-seq as previously described[32]. In brief, cells in a 10 cm dish at 80% confluency were treated in biological duplicates at the specified time points. After the specified treatment, we supplemented the treatment medium with 1 mM 4SU and metabolically labelled the cells for 10 min. The cells were lysed in QIAzol (Qiagen, 79306) and total RNA was isolated according to the manufacturer's instructions before the addition of 100 ng of RNA spike-in mix together with QIAzol. The RNA spike-in was extracted from *Drosophila* S2 cells using 4SU, metabolically labelling the cells for 20 min. The 100 µg RNA (in a total volume of 100 µl) was fragmented by addition of 20 µl of 1 M NaOH and left on ice for 20 min. Fragmentation was stopped by addition of 160 µl of 0.5 M Tris pH 6.8 and cleaned up twice using the Micro Bio-Spin P-30 Gel Columns (BioRad, 7326223) according to the manufacturer's instructions. Biotinylation of 4SU-residues was performed in a total volume of 250 µl, containing 10 mM Tris-HCl pH 7.4, 1 mM EDTA and 5 mg of MTSEA biotin-XX linker (Biotium, BT90066) for 30 min at room temperature in the dark. RNA was then purified by phenol–chloroform extraction, denatured by 10 min incubation at 65 °C and added to 200 µl µMACS Streptavidin MicroBeads (Milentyl, 130-074-101). RNA was incubated with beads for 15 min at room temperature and beads were applied to a µColumn in the magnetic field of a µMACS magnetic separator. The beads were washed twice with pull-out wash buffer (100 mM Tris-HCl, pH 7.4, 10 mM EDTA, 1 M NaCl and 0.1% Tween-20). Biotinylated RNA was eluted twice by addition of 100 mM DTT and cleaned up using the RNeasy MinElute kit (Qiagen, 74204) using 1,050 µl ethanol (≥99%) per 200 µl reaction after addition of 700 µl of RLT buffer to precipitate RNA of less than 200 nucleotides. A total of 200 ng of the purified 4SU-labelled RNA was then used as input for the TruSeq Stranded Total RNA kit (Illumina, 20020596) for library preparation. The libraries were amplified according to the manufacturer's instructions with modifications as previously described[32]. The library was amplified with 10 PCR cycles and quality-control checked on the TapeStation (Agilent) using the High Sensitivity DNA kit before pooling and paired-end sequencing on the NextSeq 550 (Illumina) system.

For DRB/TT_{chem}-seq, cells were incubated in 100 µM DRB (Sigma-Aldrich, D1916) for 3.5 h. The cells were then washed twice in PBS, and the prewarmed fresh DRB-free medium was added to restart transcription. The RNA was labelled in vivo with 1 mM 4SU for 10 min before the addition of QIAzol, which was used to stop the reaction at the desired time point.

mNET–seq

We performed mNET–seq with minor modifications to the original protocol[27]. The Flag epitope tag was added to the N terminus of the first RNAPII subunit (RPB1) in DPY30–mAID degron cells (RPB1–APEX2 cells). Cells were seeded the day before the experiment to get 100 million cells per sample the next day. We randomly assigned flasks for each treatment and treated cells with DMSO only or auxin ligand, before extracting the chromatin-bound RNAPII. The cells were first washed with ice-cold DPBS, resuspended in 4 ml of ice-cold HLB + N buffer (10 mM Tris-HCl (pH 7.5), 10 mM NaCl, 2.5 mM $MgCl_2$, 0.5% (v/v) NP-40 and 1× proteinase inhibitor) and incubated on ice for 5 min, then cells were scraped to a 15 ml centrifugate tube. The cell suspension was then underlaid with 1 ml of HLB + NS buffer (10 mM Tris-HCl pH 7.5, 10 mM NaCl, 2.5 mM $MgCl_2$, 0.5% (v/v) NP-40, 10% (w/v) sucrose and 1× proteinase inhibitor) and centrifuged to pellet the nuclei at 400*g* at 4 °C. The supernatant and membrane debris were then removed, and the nuclei were resuspended in 125 µl of NUN1 lysis buffer (20 mM Tris-HCl pH 8.0, 75 mM NaCl, 0.5 mM EDTA, 50% (v/v) glycerol and 1× proteinase inhibitor) to which we added 1.2 ml NUN2 buffer (20 mM HEPES-KOH pH 7.6, 300 mM NaCl, 0.2 mM EDTA, 7.5 mM $MgCl_2$, 1% (v/v) NP-40, 1 M urea and 1× proteinase inhibitor) to precipitate the chromatin, and the sample was incubated on ice for 15 min with occasional vortexing. The lysates were then centrifuged at 16,000*g* for 10 min at 4 °C to pellet the chromatin. The chromatin pellets were then washed with 1× MNase buffer and then digested with 50 U MNase (NEB, M0247S) for 2 min at 37 °C with 1,400 rpm on a thermomixer. The digestion was stopped by adding 5 µl of 500 mM EGTA (to a final concentration of 25 mM; Thermo Fisher Scientific, 50-255-956) and transferred onto ice. The reactions were then centrifuged at 16,000*g* for 5 min at 4 °C and the supernatant was subsequently diluted with 1 ml of NET-2 buffer (50 mM Tris-HCl pH 7.4, 150 mM NaCl, 0.05% (v/v) NP-40) per fraction and pooled per sample for the N-terminal Flag-RNAPII IP. We added 50 µl of anti-Flag M2 Affinity gel (Sigma-Aldrich, A2220) and incubated in the cold room for 1 h. This was followed by eight washes with NET-2 buffer and one wash with 500 µl of PNKT buffer containing 1× T4 polynucleotide kinase (PNK) buffer (NEB, M0201L) and 0.1% (v/v) Tween-20 (Thermo Fisher Scientific, BP337-100). The beads were incubated in 100 µl of PNK reaction mix containing 1× PNK buffer, 0.1% (v/v) Tween-20, 1 mM ATP and T4 PNK 3′ phosphatase minus (NEB, M0236L) at 37 °C for 10 min. We eluted the RNA by adding 350 µl of buffer RLT plus and 1 ng of fragmented *Drosophila* S2 mRNA spike-in to the extraction buffer. Next, short immunoprecipitated RNA fragments were size-selected (under 200 nucleotides) and purified from the eluates according to the manufacturer's protocol after the purification of miRNA from animal cells using the RNeasy Plus Mini Kit and the RNeasy MinElute Cleanup Kit (Qiagen, 74204), and eluted in 14 µl of nuclease-free water. We checked 1 µl of the eluted RNA on the TapeStation for the RNA

quality and proceeded to NGS library preparation using the NEBNext Multiplex Small RNA Library Prep kit (NEB, NC0477293). Both were performed according to the manufacturer's protocol with low input material, with 14 PCR cycles for the library amplification step. We cleaned up and concentrated the DNA using the Monarch kit (NEB) and separated the library on a 6% TBE gel and performed size-selection by excising the smear between 147 and 307 nucleotides (according to the Quick-Load pBR322 DNA-MspI Digest ladder (NEB)). These libraries were quality-checked and quantified using the TapeStation. The samples were pooled and sequenced using paired-end sequencing on the NextSeq 550 (Illumina) system.

For promoter-proximal RNAPII half-life experiments determined by mNET–seq, chromatin was isolated from cells that were pretreated with DMSO or auxin and then incubated with 1 μM triptolide for 0, 5, 10, 20 or 40 min in the presence of DMSO/auxin and analysed using mNET–seq. Chromatin was digested with micrococcal nuclease (MNase, scissor) to release RNAPII engaged RNA from insoluble chromatin, and then immunoprecipitated using anti-Flag–RNAPII antibodies.

IP–MS data analysis

Data were analysed in Proteome Discoverer v.3.1. A database search was performed with the Sequest HT search engine using the Mouse UniProt database containing only reviewed entries and canonical isoforms (retrieved on 10 October 2019). Oxidation (M) was set as a variable modification, while TMT was set as fixed modification. A maximum of two missed cleavages were permitted. The precursor and fragment mass tolerances were 10 ppm and 0.6 Da, respectively. Peptide-spectrum matches (PSMs) were validated by percolator with a 0.01 posterior error probability threshold. Only PSMs with an isolation interference of <25% and at least 5 MS2 fragments matched to the peptide sequence selected for MS3 were considered. The quantification results of PSMs were combined into protein-level quantification using the MSstatsTMT R package[56]. Only proteins with at least three peptides were reported. To identify interactors, we performed differential abundance analysis between the IP samples and their corresponding controls (that is, 0 h IP was compared to 0 h IgG control and 8 h IP was compared to 8 h IgG control). A protein was considered to be an interactor if in one or both comparisons its levels were statistically significantly different ($Q \leq 0.05$, limma test, with P values adjusted by the Storey method) and at least twice higher in IP reactions than in the corresponding IG control (Supplementary Table 3).

SILAC/MS analyses

All SILAC/MS data were processed using the MaxQuant software (Max Planck Institute of Biochemistry; v.1.5.3.30). The default values were used for the first search tolerance and main search tolerance—20 ppm and 6 ppm, respectively. Labels were set to Arg10 and Lys8. MaxQuant was set up to search the reference mouse proteome database downloaded from UniProt on 9 January 2020. MaxQuant performed the search assuming trypsin digestion with up to two missed cleavages. Peptide, site and protein false-discovery rate (FDR) were all set to 1% with a minimum of one peptide needed for identification but two peptides needed to calculate a protein level ratio. The following modifications were used as variable modifications for identifications and included for protein quantification: oxidation of methionine, acetylation of the protein N terminus, ubiquitination of lysine, phosphorylation of serine, threonine and tyrosine residues, and carbamidomethyl on cystine. Intensity values measured in all replicates were $\log_2$-transformed (Supplementary Table 4), P values were computed using Fisher's tests and corrected using Benjamini–Hochberg FDR correction. All raw MS data files have been deposited to the ProteomeXchange Consortium (and PXD039176). The Gene Ontology term enrichment analysis was performed using Enrichr online tool (https://maayanlab.cloud/Enrichr/), STRING (https://string-db.org) and clusterProfiler[57].

ChIP–seq data analysis

The sequenced reads were demultiplexed using bcl2fastq (v.2.19.0.316), and basic quality control was performed on the resulting FASTQ files using FastQC (v.0.11.8). FASTQ reads were mapped to the GRCm38 (mm10) genome using Bowtie2 (v.2.4.1) using the standard settings. The resulting SAM files were converted to BAM files using the SAMtools (v.1.10) view command, after which the BAM files were sorted and indexed, and potential PCR duplicates were removed using the rmdup function. DeepTools (v.3.3.0) was used to generate occupancy heat maps, and the resulting normalized occupancy matrix was used as input for public R scripts to generate average profile plots and to calculate processivity indices. In brief, the BAM files were converted into BigWig files using the bamCoverage function (bamCoverage -p 8 --normalizeUsing RPGC --effectiveGenomeSize mm10 --centerReads -e --scaleFactor X --blackListFileName mm10.blacklist.bed). For comparison, quantitative ChIP–seq data using spike-in normalization were used, normalization to the *Drosophila* S2 spike-in was performed at this stage according to the manufacturer's instructions (Active Motif, 61686; https://www.activemotif.com/catalog/1091/chip-normalization). The computeMatrix function was used to quantify the occupancy of reads across the specified intervals, and the plotProfile and plotHeatmap functions were used to plot the data. Reproducibility of replicates is shown in Supplementary Fig. 3.

Quant-seq/SLAM-seq analysis

Gene and 3′ untranslated region (UTR) annotations were obtained from the UCSC table browser (https://genome.ucsc.edu/cgi-bin/hgTables, mm10 vM14 3′ UTR). Adapters were trimmed from raw reads using cutadapt through the trim_galore wrapper tool with adapter overlaps set to 3 bp for trimming. For Quant-seq, concatenated fastq files were trimmed for adapter sequences, and masked for low-complexity or low-quality sequences using trim_galore, then mapped to the mm10 whole genome using HISAT v.2.2.1 with the default parameters. The number of reads mapped to the 3′ UTR of genes was determined using featureCounts. Raw reads were normalized to CPM. SLAM-seq analysis was performed as previously described[58] using the SlamDunk package[59]. Trimmed reads were further processed with SlamDunk (v.0.3.4 16). The 'Slamdunk all' command was executed with the default parameters except '-rl 74 -t 8 fastq.gz -n 100 -m -mv 0.2 -o Slamdunk2', running the full analysis procedure (slamdunk all) and aligning against the mouse genome (GRCm38), filtering for variants with a variant fraction of 0.2. Unless indicated otherwise, reads were filtered for having ≥2 T>C conversions. The remaining parameters were left as defaults.

Analysis of differential gene expression was restricted to genes with ≥10 reads in at least one condition. Differential gene expression calling was performed on raw read counts with ≥2 T>C conversions using DESeq2 with the default settings, and with size factors estimated on corresponding total mRNA reads for global normalization. Downstream analysis was restricted to genes that passed all internal filters for FDR estimation by DESeq2. Plots of differential gene expression were visualized using the ggplot2 package in R with significant genes (P value < 0.05, $|\log2FC| \geq 1$). Reproducibility of replicates is shown in Supplementary Fig. 3.

mNET–seq data analysis

Reads were demultiplexed using bcl2fastq, then trimmed for adapter content with cutadapt (-m 10 -e 0.05 --match-read-wildcards -n 1), and mapped with STAR to the GRCm38 (mm10) genome assembly. Further data processing was performed using the R/Bioconductor environment. Coverage tracks for further analysis were restricted to the last nucleotide incorporated by the RNAPII in the aligned mNET–seq reads as described[60]. To calculate the half-life of paused RNAPII for each gene by mNET–seq, RNAPII density was calculated in a 300 bp window downstream of the TSS. RNAPII time-course measurements were fitted into

an exponential decay model using the RNAdecay R package (https://bioconductor.org/packages/release/bioc/html/RNAdecay.html). We selected genes fulfilling the current criteria: (1) detectable RNAPII levels (reads per kilobase of transcript, per million mapped reads > 1), (2) highest RNAPII density under the no triptolide (0 min) condition and (3) low variance between replicates ($\sigma < 0.05$). Genes fitting the above criteria ($n = 6,338$) were used to calculate the RNAPII half-life. Reproducibility of replicates is shown in Supplementary Fig. 3.

TT_{chem}-seq data analysis

Paired-end reads were demultiplexed using bcl2fastq. TT_{chem}-seq raw data were processed essentially as described previously[32]. Raw reads were aligned to the mouse mm10 genome assembly using STAR. Mapped reads with a mapping quality score <10 were discarded with SAMtools. All further processing was performed using the R/Bioconductor framework. Antisense bias, sequencing depth and cross-contamination rates were calculated as described previously[32]. Reads were mapped to transcription units, which represent the union of all annotated UCSC RefSeq isoforms per gene. The number of transcribed bases per transcription unit was calculated as the sum of the coverage of evident (sequenced) fragment parts (read pairs only) for all fragments in addition to the sum of the coverage of the inner mate interval if not entirely overlapping a RefSeq annotated intron (UCSC RefSeq GRCm38). Computational analysis DRB/TT_{chem}-seq data were processed using a previously published protocol[32]. In brief, reads were aligned to human GRCm38 (mm10). Read depth coverage was normalized to account for differences between samples using a scale factor derived from a spike-in aligned and counted against *Drosophila melanogaster*. Biological replicate alignments were combined for the purpose of visualization and wave-peak analysis to increase read-depth coverage.

A set of non-overlapping protein-coding genes of >60 kb and <300 kb was selected for wave-peak analysis. A meta-gene profile was calculated by taking a trimmed mean of each base-pair coverage in the region around the TSS. This was further smoothened using a spline. Wave peaks were called at the maximum points on the spline, with the stipulation that the peak must advance with time before being subjected to manual review. Elongation rates (kb per min) were calculated by fitting a linear model to the wave-peak positions as a function of time. For elongation-rate analysis, the following criteria were used to filter genes: The 0 min timepoint DMSO control sample was required to show expression of the gene (mean expression of >100 rpm by TT-seq) and was required to have a wave peak called within 10 kb of the pausing peak region to remove artifacts. Genes showing an increase in transcription in the DMSO control sample for the time course were identified by requiring the wave peak in the 0 min sample to be less than the wave peak in the 10 min timepoint wave peak, and the wave peak in the 10 min sample to be less than the wave peak in the 20 min timepoint, and the wave peak in the 20 min sample to be less than the wave peak in the 30 min timepoint. This resulted in the identification of 855 genes, for which elongation rates were calculated for the samples by dividing the wave peak position by the timepoint. Reproducibility of replicates is shown in Supplementary Fig. 3.

Statistics and reproducibility

The statistical details of the experiments can be found in the figure legends and in the Methods. Western blotting in Figs. 1b,c,f–h, 3a and 5b,c,j was independently performed three times with similar results, and western blotting in Extended Data Figs. 8c,f,g and 9c,e was performed twice as biologically independent experiments.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All related raw sequencing and processed data have been deposited at the Gene Expression Omnibus under accession number GSE181714.

53. Spruijt, C. G., Baymaz, H. I. & Vermeulen, M. Identifying specific protein–DNA interactions using SILAC-based quantitative proteomics. *Gene Reg.* **977**, 137–157 (2013).
54. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V. & Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **1**, 2856–2860 (2006).
55. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).
56. Choi, M. et al. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* **30**, 2524–2526 (2014).
57. Wu, T. et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141 (2021).
58. Muhar, M. et al. SLAM-seq defines direct gene-regulatory functions of the BRD4-MYC axis. *Science* **360**, 800–805 (2018).
59. Neumann, T. et al. Quantification of experimentally induced nucleotide conversions in high-throughput sequencing datasets. *BMC Bioinform.* **20**, 258 (2019).
60. Prudencio, P., Rebelo, K., Grosso, A. R., Martinho, R. G. & Carmo-Fonseca, M. Analysis of mammalian native elongating transcript sequencing (mNET-seq) high-throughput data. *Methods* **178**, 89–95 (2020).

Acknowledgements We thank H. Damhofer, K. Nishimura, D. Shlyueva, S. Sidoli, Z. Li, Z. Sun, Y. Lin and Y. Fang for technical advice and reagents; C. Huang, A. Radzisheuskaya, D. Shlyueva and R. Armstrong for comments on the manuscript; and members of the Helin laboratory for discussions and support. This work was supported by the Memorial Sloan Kettering Cancer Center Support Grant (no. NIH P30 CA008748), a Tri-Institutional Stem Cell grant (no. 2019-035), startup funds from The Institute of Cancer Research and the Novo Nordisk Foundation to the NNF Center for Stem Cell Biology (no. NNF17CC0027852).

Author contributions H.W. and K.H. conceived the project. H.W. designed and performed the experiments and bioinformatics analysis. P.V.S. and M.M. performed the MS studies, supervised by R.C.H. Z.F. and X.J. provided input to the project. H.W. and K.H. wrote the manuscript. K.H. supervised the study and acquired funding.

Competing interests K.H. is a co-founder of Dania Therapeutics, consultant for Inthera Bioscience and a scientific advisor for MetaboMed and Hannibal Innovation. The other authors declare no competing interests.

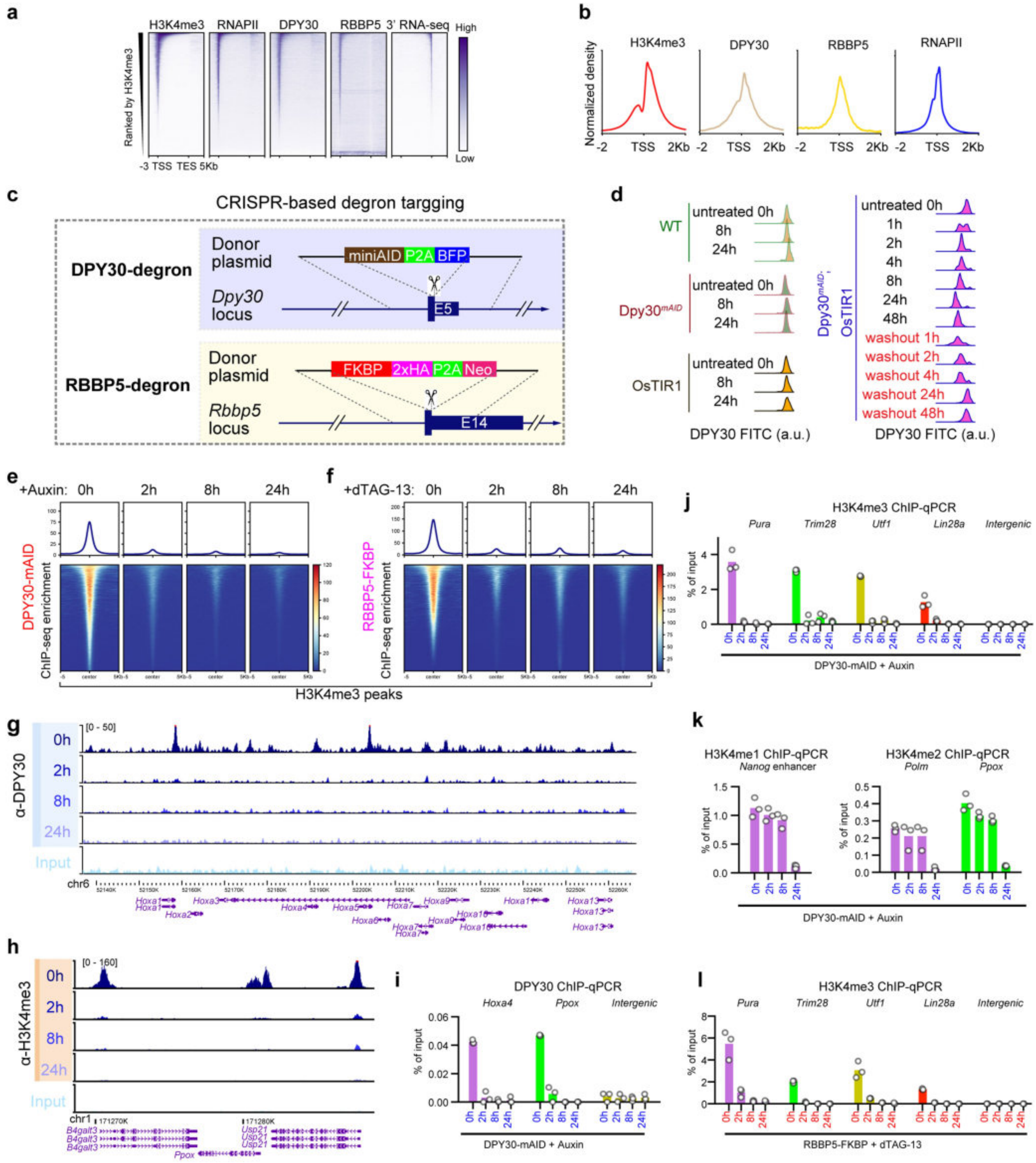
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-05780-8.

Correspondence and requests for materials should be addressed to Kristian Helin.

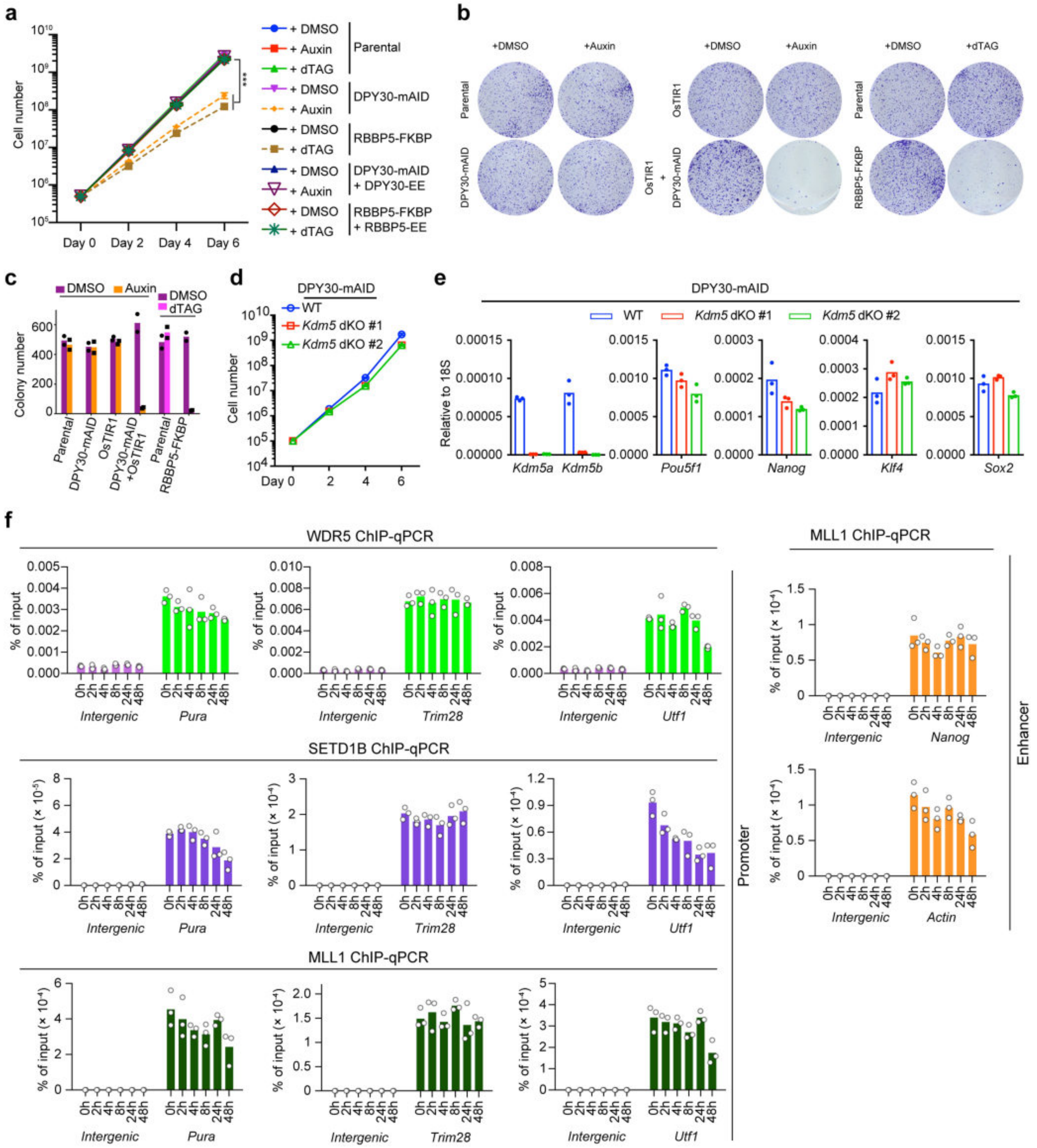
Peer review information *Nature* thanks Alvaro Rada-Iglesias and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at http://www.nature.com/reprints.

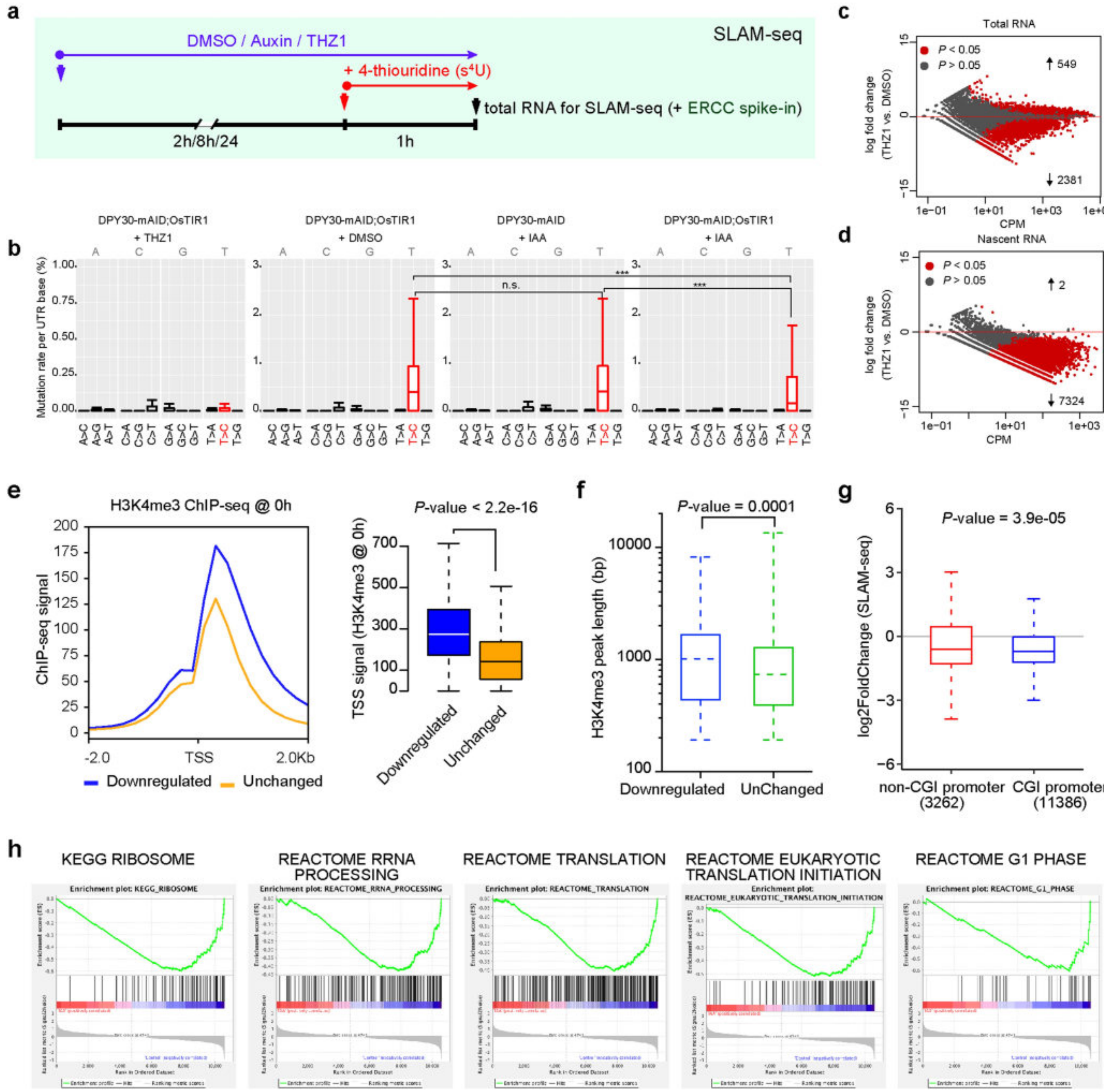


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Generation of DPY30–mAID and RBBP5–FKBP mES cell lines. (**a**,**b**) DPY30, RBBP5, H3K4me3 and RNAPII occupancy in mES cells plotted around the TSS of all protein-coding genes. Genes were sorted based on H3K4me3 binding levels in the heat maps (**a**). 3′ RNA-seq data is from Quant-seq in mES cells. The average signal in the profiles (**b**) is plotted over the transcription start sites (TSS ± 2 kb) of all protein-coding genes. (**c**) Outline of CRISPR-HDR based knock-in targeting approach generating DPY30-AID and RBBP5–FKBP, by knocking-in mAID and FKBP12 degron tags into the 3′ end of both alleles of endogenous *Dpy30* and *Rbbp5* loci of the E14 mES cell cell line, respectively. The endogenous *Dpy30* gene was edited to encode Auxin inducible degradation (mAID)-BFP at the C terminus of DPY30 in OsTir1 E14 mES cells. The endogenous *Rbbp5* gene was edited to encode FKBP12 tag-Neomycin that enables targeted degradation upon dTAG-13 treatment. (**d**) DPY30 expression in the indicated cell lines as measured by flow cytometry analysis using an anti-DPY30 antibody. (**e,f**) H3K4me3 ChIP–seq heat maps and profiles at H3K4me3 peak centre in control and Auxin-treated cells in DPY30–mAID (E) or dTAG-13-treated in RBBP5–FKBP (F) cells. The signal was plotted over H3K4me3 peaks centre (peak centre ± 5 kb) which are called from WT mES cells. Sites are sorted by the ChIP–seq signals at 0 h. (**g,h**) Integrative genomics viewer (IGV) browser snapshots comparing DPY30 (g) or H3K4me3 (h) enrichments determined by ChIP–seq in control and Auxin-treated in DPY30–mAID degron mES cells. (**i**–**k**) ChIP-qPCR enrichments of DPY30 (I) and H3K4 methylations (j and k) in DPY30–mAID cells after treatment with Auxin for the indicated times. (**l**) ChIP-qPCR for H3K4me3 in RBBP5–FKBP cells. In panels (i-l) target sites around the promoter of the genes and control region (intragenic chr8: 72,806,101- 72,806,240) were used. For H3K4me1, the enhancer region of *Nanog* was used for DPY30, H3K4me2 and H3K4me3 ChIP-qPCR. Graph shows mean values from technical triplicates (n = 3), from one representative out of two independent experiments.

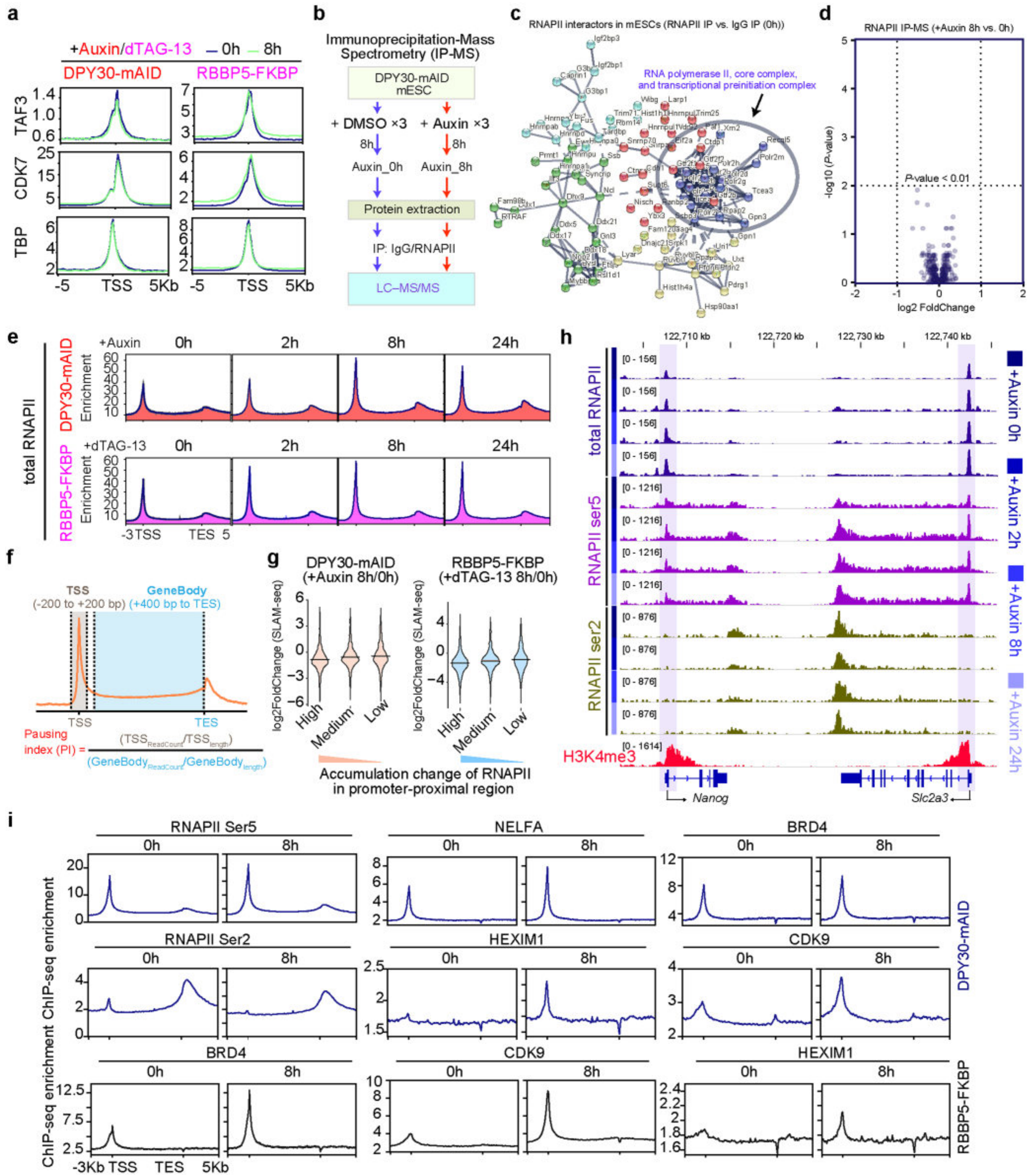


Extended Data Fig. 2 | DPY30 and RBBP5 are required for cell proliferation. (**a**) Cell proliferation assays of DPY30–mAID and RBBP5–FKBP cells grown either with or without Auxin/dTAG-13. The line graph represents the mean ± SD of the numbers of cells counted at each time point. Parental represents the parental E14 ES cells without CRISPR editing. Ectopic expression (EE) of DPY30 or RBBP5 rescues the proliferation of defect observed by degrading endogenous DPY30 and RBBP5, respectively. Data are from three biological replicates (n = 3) and were analysed using Two-way ANOVA (n = 3) and represented as mean ± s.d., ****P* < 0.001. (**b**) Colony formation assay for DPY30–mAID and RBBP5–FKBP cells grown either with or without Auxin/dTAG-13 for two weeks. (**c**) Quantification of colony formation assays from two independent experiments. Data in bar plots are represented as mean and n = 2 replicates. (**d**) Cell proliferation assays of DPY30–mAID cells after treatment with Auxin at the indicated times with or without *Kdm5ab* dKO. (**e**) RT-qPCR analyses of mRNA expression of DPY30–mAID cells with or without *Kdm5ab* dKO. Two independent dKO clones were chosen for downstream analysis. The values are normalized to 18S rRNA. Graph shows mean values from technical triplicates (n = 3), from one representative out of two independent experiments. (**f**) ChIP-qPCR enrichment for WDR5, SETD1B and MLL1 in DPY30–mAID cells system after treatment with Auxin at the indicated points. Graph shows mean values from technical triplicates (n = 3), from one representative out of two independent experiments.



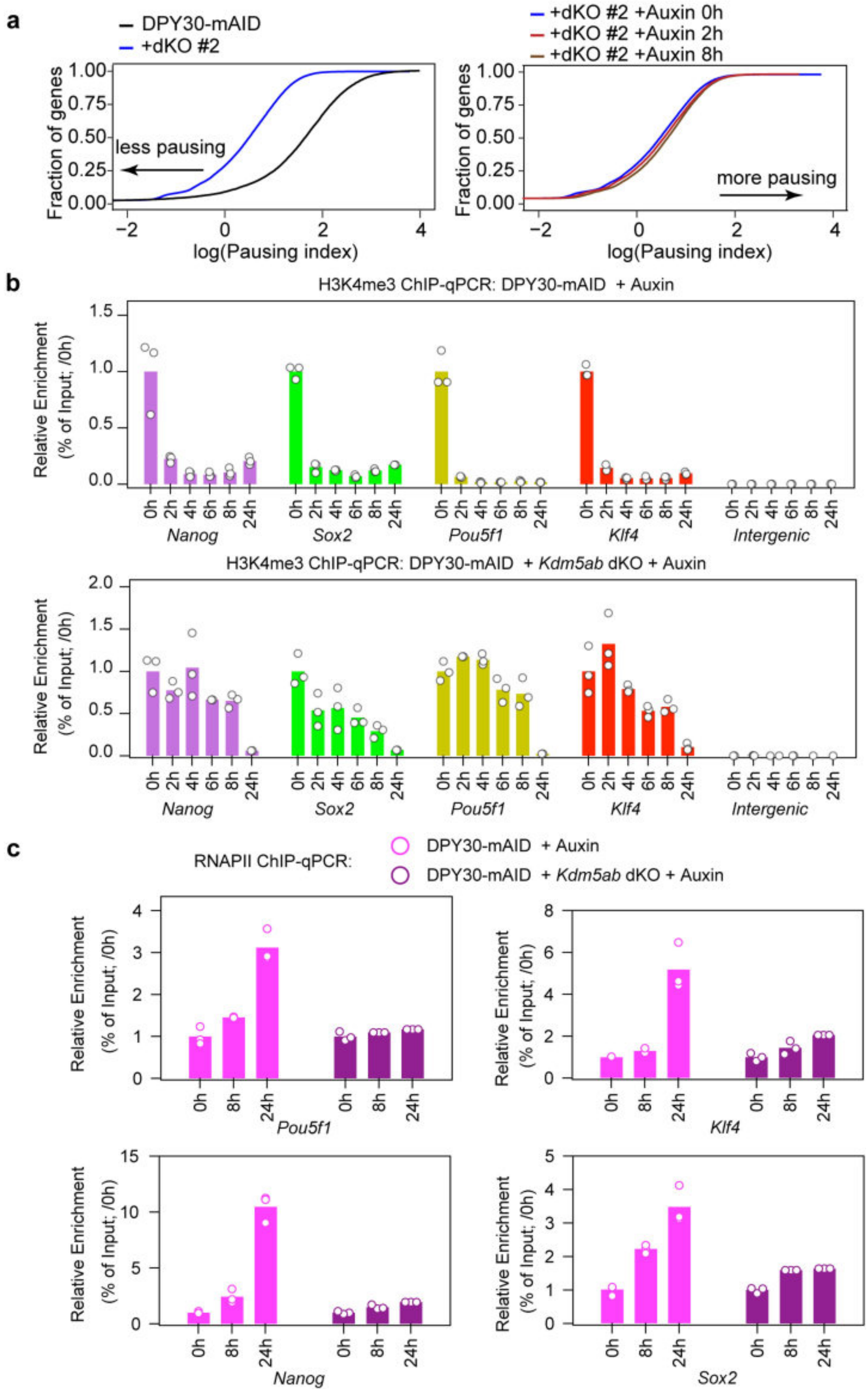
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Outline and controls for SLAM-seq experiments. (**a**) Experimental design. To validate the SLAM-seq protocol, we performed a pilot experiment for mapping responses to short-term THZ1 (2h) treatment by SLAM-seq in mES cells. SLAM-seq utilizes thymine-to-cytosine (T>C) conversion from 4-thiouridine (4sU)-labelled mRNAs to quantify the abundance of nascent RNA transcripts using 3′ end mRNA-sequencing (Quant-seq). To monitor the consistency and reproducibility of different SLAM-seq data, we inhibited transcription with THZ1 (reduces RNAPII-mediated gene transcription by inhibiting cyclin-dependent kinase 7 (CDK7)). (**b**) Conversion rates for each position of 4-thioU-containing SLAM-seq reads (≥ 2 T>C conversions) before or after Auxin or THZ1 treatment for 2 h. Changes in the abundance of newly synthesized mRNAs (detected in SLAM-seq based on T>C conversions). Average conversion rates (centre line) ± s.d. (whiskers) of two independent experiments (points) are shown. *P* value (Two-sided Mann-Whitney test) is indicated (****P* < 0.001, n.s., not significant.). n = 20,428 transcripts. The boxplot indicates the median (middle line) and the third and first quartiles (box); the whiskers show the 1.5× IQR above and below the box. (**c,d**) Transcriptional response of the cells treated with THZ1/DMSO for 2h followed by 4sU labelling over 60 min. (c) MA plots comparing total gene expression level with log change in transcription per gene measured by 4-thioU RNA-seq (SLAM-seq). (d) MA plots comparing nascent gene expression levels with log change in transcription per gene measured by SLAM-seq. THZ1 treatment confirmed that transcripts containing T>C conversions of protein-coding genes were broadly repressed, which captured the prominent immediate responses, while the total mRNA level showed fewer changes. *P*-adjusted value by Wald test in DESeq2. (**e**) H3K4me3 levels on TSS before Auxin treatment of downregulated genes (n = 1,111) and unchanged genes (n = 10,107) measured by SLAM-seq (in response to Auxin treatment for 2h in DPY30–mAID cells). The *p* value was calculated with a two-sided Wilcoxon test. The boxplot indicates the median (middle line) and the third and first quartiles (box); the whiskers show the 1.5× IQR above and below the box. (**f**) H3K4me3 peak width of downregulated genes and unchanged genes measured by ChIP–seq at steady-state. n= for downregulated and n= for unchanged genes. The *p* value was calculated with a two-sided Wilcoxon test. The boxplot indicates the median (middle line) and the third and first quartiles (box); the whiskers show the 1.5× IQR above and below the box. (**g**) Box plot showing log2-transformed fold change of nascent transcription (SLAM-seq, 2 h vs. 0 h) of genes containing CGI (n = 11,386) and non-CGI (n = 3,262) promoters. The *p* value was calculated with a two-sided Wilcoxon test. The boxplot indicates the median (middle line) and the third and first quartiles (box); the whiskers show the 1.5× IQR above and below the box. (**h**) Gene set enrichment analysis (GSEA) of downregulated genes in DPY30–mAID cells in response to 2 h Auxin treatment.

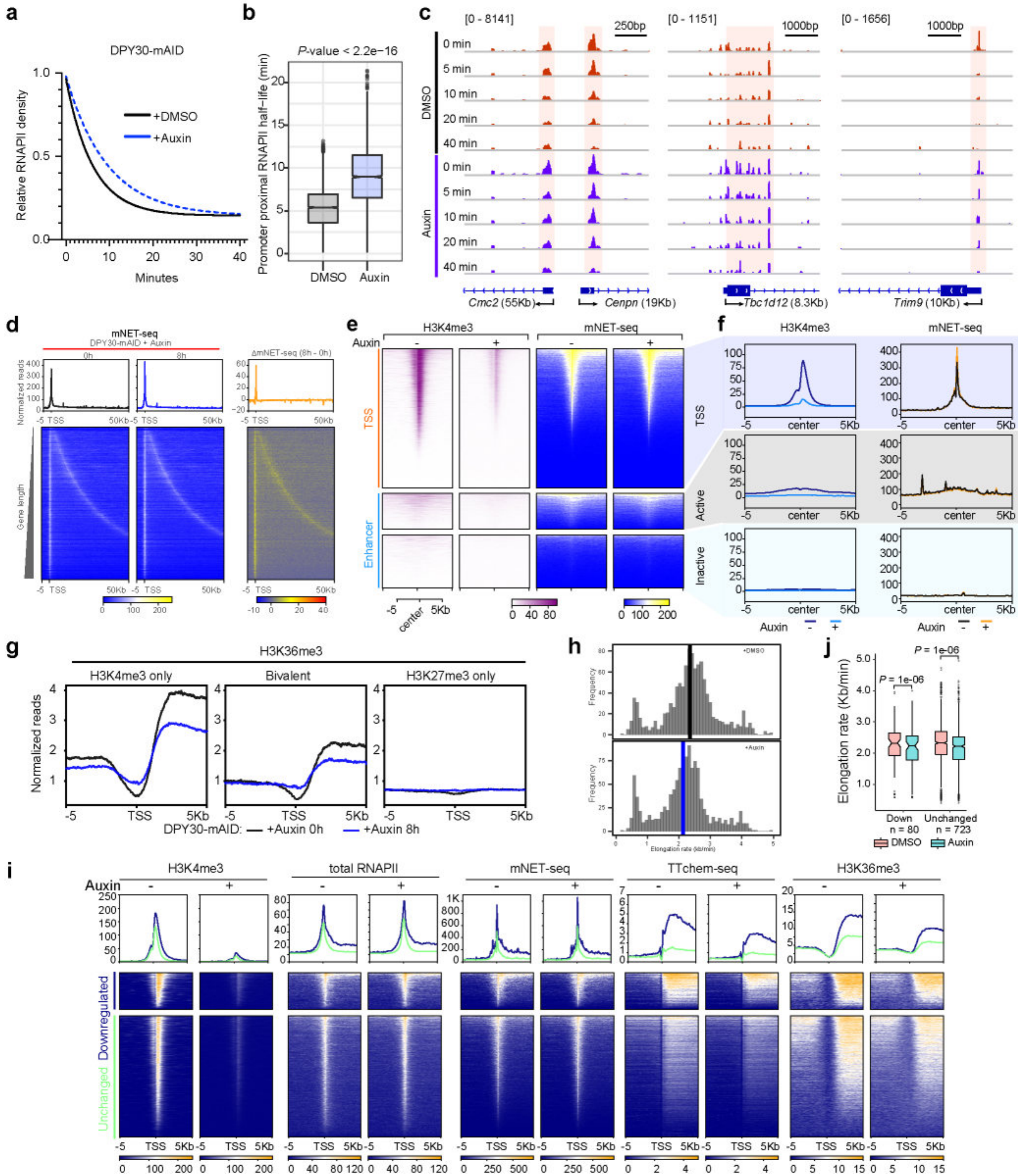


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | H3K4me3 loss does not have detectable effects on binding of TAF3, CDK7 and TBP to transcription start sites and to PIC formation. (**a**) ChIP–seq profiles of the indicated proteins using DPY30–mAID and RBBP5–FKBP cells treated with or without Auxin/dTAG-13, respectively for 8 h. The enrichments were plotted over the transcription start sites (TSS ± 5 kb) of protein-coding genes. TSS, transcription start site. (**b**) Outline of the mass spectroscopy proteome profiling strategy for mapping the interaction networks of RNAPII in DPY30–mAID cells treated with or without Auxin. (**c**) String network of protein complexes (*k*-means clustering) showing RNAPII interactors in control cells compared with IgG mock IP. (**d**) Volcano plot showing proteins changing their association with RNAPII in response to acute loss of H3K4me3. The *x* axis displays the enrichment (log2 fold change) of proteins in Auxin-treated cells (Auxin 8 h) compared to DMSO-treated control cells (Auxin 0 h). The *y* axis shows the significance (-log10 *P* value) of enrichment calculated from three biological replicate experiments. A protein was considered an interactor if in one or both comparisons its levels were statistically significantly different (Q value ≤ 0.05, limma test, with *P* values adjusted by the Storey method). (**e**) RNAPII occupancy based on ChIP–seq in the indicated cell lines. Metagene analysis showing the genome-wide enrichment averages on protein-coding genes, data are shown along with 3 kb upstream of the transcriptional start site to 5 kb downstream of the end of each annotated gene. TSS, transcription start site, TES, transcription end site. (**f**) Estimation of a gene's "pausing index" (PI) from RNAPII ChIP–seq data. The promoter is defined as the region covering 200 bp upstream to 200 bp downstream of the TSS; the gene body is defined as the region from 400 bp downstream of the TSS to TES, genes with gene length less than 400 bp are removed for pausing index analysis. (**g**) Violin plots showing changes of gene expression in the indicated samples. Genes were separated into three equal parts based on their accumulation change of RNAPII in promoter-proximal region. (**h**) An IGV snapshot comparing RNAPII, RNAPII Ser 2p and Ser 5p ChIP–seq signals in control and Auxin-treated DPY30–mAID cells at the indicated times. (**i**) Average metagene ChIP–seq profiles for the indicated factors in control and dTAG-13-treated RBBP5–FKBP cells.

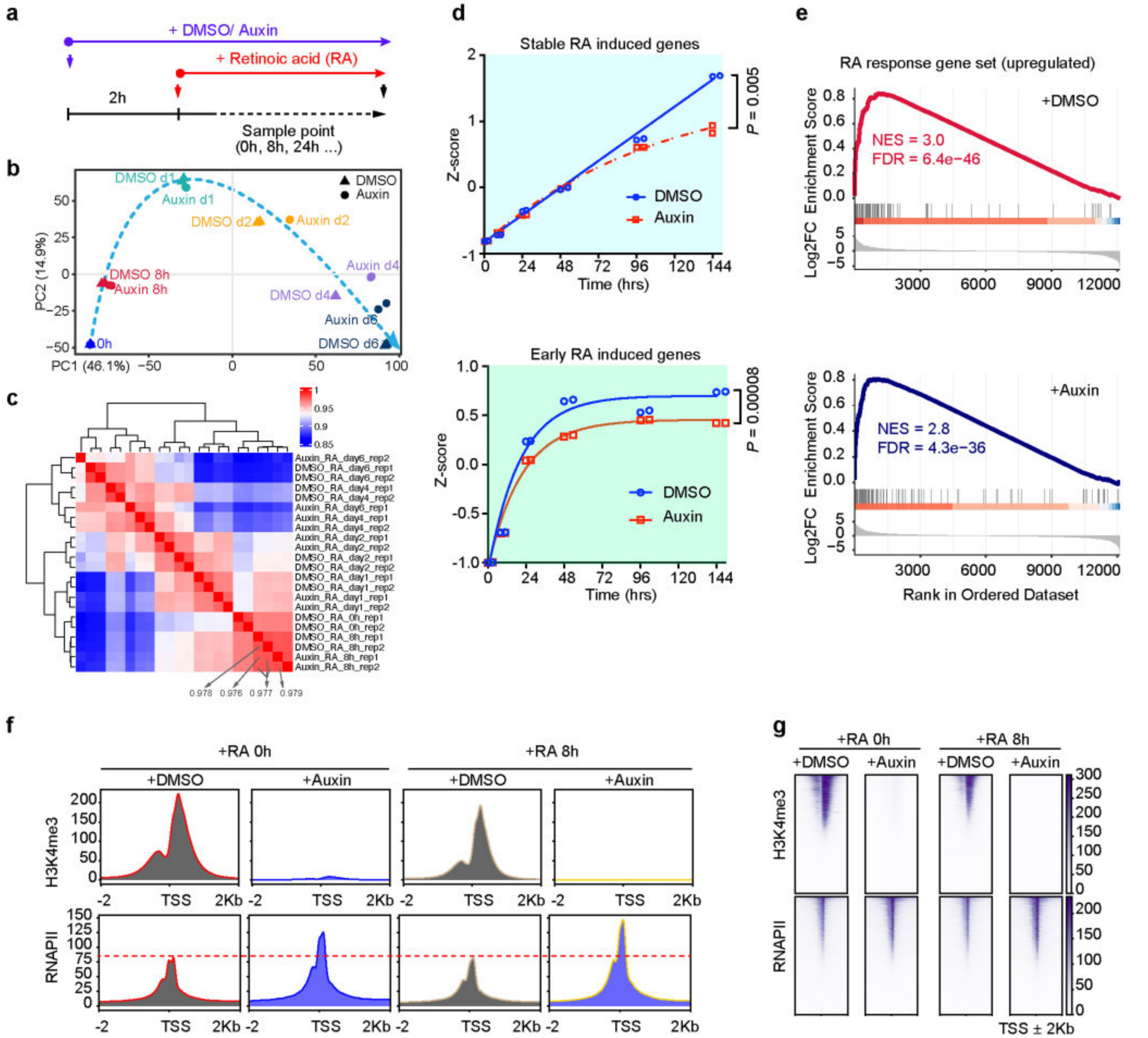


Extended Data Fig. 5 | The fast turnover of H3K4me3 is dependent on KDM5 demethylases. (**a**) RNAPII pausing index in DPY30–mAID (black), DPY30–mAID: *Kdm5a/b* dKO (blue, clone #2) and Auxin-treated cells. Higher index values indicate a higher degree of RNAPII pausing. (**b**) ChIP-qPCR enrichment of H3K4me3 in DPY30–mAID cells after treatment with Auxin at the indicated times with or without *Kdm5a/b* dKO. (**c**) ChIP-qPCR signals for RNAPII in DPY30–mAID cells following treatment with Auxin at the indicated times with or without *Kdm5a/b* dKO. For ChIP-qPCR, the target sites around promoter of the indicated genes and control region (intragenic chr8: 72,806,101- 72,806,240) were used. Graph shows mean values from technical triplicates (n = 3), from one representative out of two independent experiments.

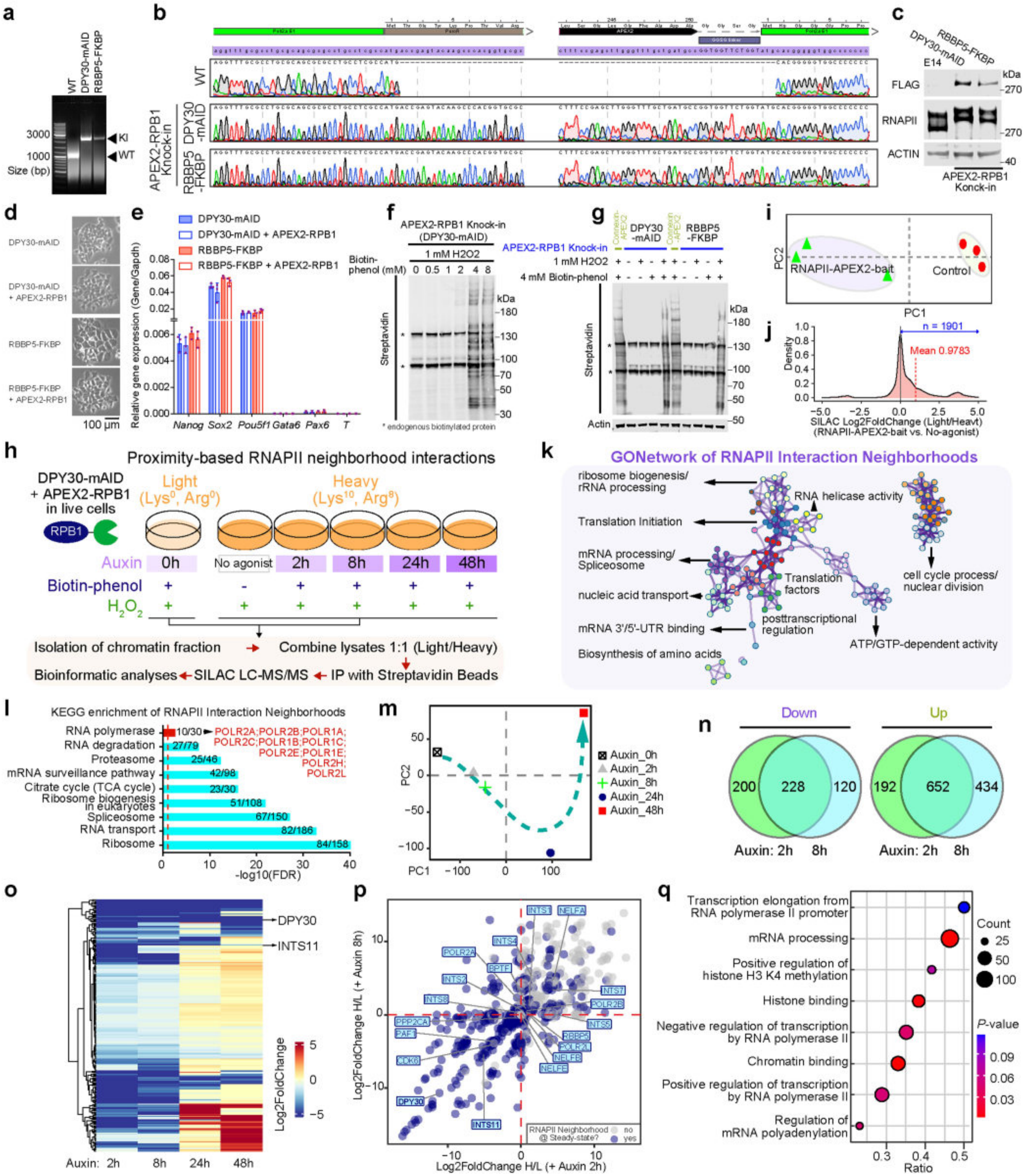


Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | H3K4me3 regulates the paused RNAPII half-life. (**a**) Half-lives of paused RNAPII in control and Auxin-treated DPY30–mAID degron cells. The half-life was calculated based on an exponential decay model. Normalized promoter-proximal RNAPII density for each gene is shown over the course of triptolide treatment both for control (DMSO, black) and Auxin (blue) conditions. n = 2 biological replicates. (**b**) Boxplot showing increased paused RNAPII duration following H3K4me3 loss. *P* value was calculated with a two-sided Wilcoxon test, n = 4,007 genes. The boxplot indicates the median (middle line) and the third and first quartiles (box); the whiskers show the 1.5× IQR above and below the box. (**c**) An IGV snapshot comparing mNET–seq signals after triptolide-induced block of transcription in control and Auxin-treated in DPY30–mAID cells at the indicated time points. The decreasing levels of paused RNAPII were observed in the pausing window (shaded areas) over time. (**d**) Heat maps and profiles showing changes in mNET–seq signals upon acute loss of H3K4me3. Heat maps were plotted as increasing gene length over regions 5 kb upstream to 50 kb downstream of the TSS. The right panel display the difference (Δ) in mNET–seq between control (+ Auxin 0 h) and Auxin-treated (+ Auxin 8 h) cells. Genes were sorted by gene length. (**e**–**f**) Heat maps and profiles showing changes of H3K4me3 ChIP–seq and mNET–seq at promoters, active and inactive enhancers upon Auxin treatment (8 h) in DPY30–mAID degron cells. The signals were plotted over the transcription start sites (TSS ± 5 kb) or the centre of enhancers (centre ± 5 kb). (**g**) H3K36me3 ChIP–seq profiles in control (+ Auxin 0 h) and Auxin-treated (+ Auxin 8 h) cells in DPY30–mAID cells. Genes were split by their H3K4me3 and H3K27me3 levels around TSS regions. (**h**) Histogram of RNAPII elongation rates for individual genes between 60 and 300 kb with RPM value ≥100 across all time points (n = 855) with a 10-min wave peak called beyond 2 kb and sequential increase from the TSS over the 10-, 20- and 30-min time points. (**i**) Heat maps and profiles showing changes of ChIP–seq, mNET–seq and TT_{chem}-seq at promoters upon Auxin treatment (8 h) in DPY30–mAID degron cells. The downregulated genes and unchanged genes measured by SLAM-seq. The signals were plotted over the transcription start sites (TSS ± 5 kb). (**j**) Boxplot showing RNAPII elongation rates following H3K4me3 loss. *P* value was calculated with a two-sided Wilcoxon test. The number of genes (n) from each group are shown at the bottom. The boxplot indicates the median (middle line) and the third and first quartiles (box); the whiskers show the 1.5× IQR above and below the box.

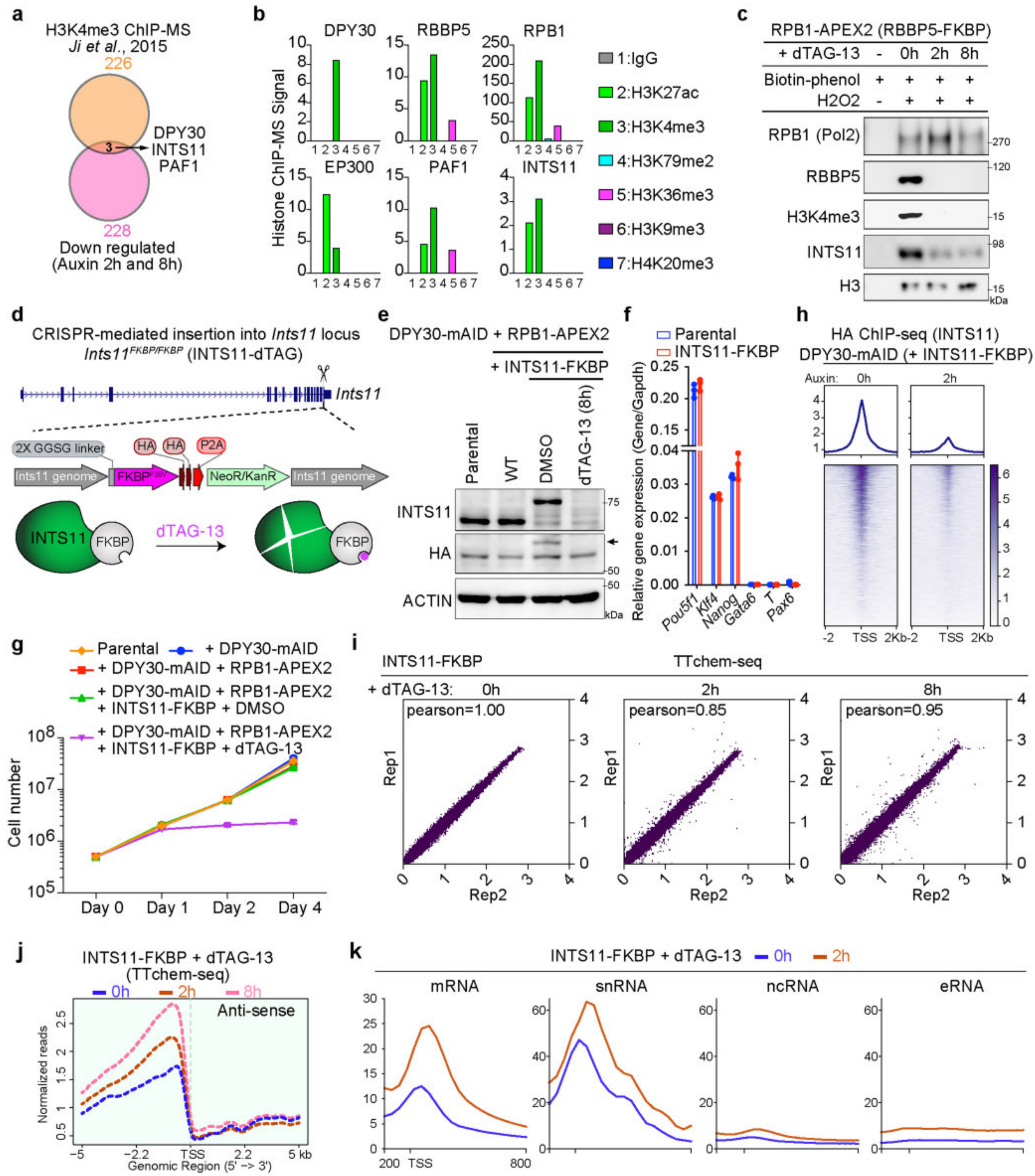


Extended Data Fig. 7 | Loss of H3K4me3 does not have detectable effects on transcriptional initiation. (**a**) Experimental strategy for retinoic acid (RA) differentiation in control and Auxin-treated degron cells. (**b**) Principal component (PC) analysis of RA-induced mRNA expression changes in DPY30–mAID cells prior to treatment with Auxin (circles) or DMSO (triangles) for 2 h at the indicated differentiation points: 0h, 8h, d1, d2, d4, d6. Developmental trajectory is shown by the dashed arrow. (**c**) Spearman correlation heat map of retinoic acid (RA) time course RNA-seq replicates of Auxin-treated and DMSO-treated cells. (**d**) Short- and long-term effects of DPY30 degradation on the expression of RA-induced genes. DPY30–mAID cells were treated with DMSO, Auxin and RA as indicated. The stable RA induced genes were identified from upregulated genes at day 6 in control cells (DMSO). The early RA induced genes were identified from upregulated genes at 8 h in control cells (DMSO). Data were analysed using Two-way ANOVA (n = 2) and represented as mean ± s.d., $^{**}P < 0.01$, $^{***}P < 0.001$. (**e**) Gene set enrichment analysis (GSEA) analysis of RA induced genes (RA upregulated gene list from previously published data, n = 227). The curve represents the evolution of the density of the genes identified in the RNA-seq. The False Discovery Rate (FDR) is calculated by comparing the actual data with 1000 Monte-Carlo simulations. The NES (Normalized Enrichment Score) computes the density of modified genes in the dataset with the random expectancies, normalized by the number of genes found in the given gene cluster, considering the size of the cluster. (**f,g**) H3K4me3 and RNAPII occupancy before or after RA treatment in control and Auxin-treated degron cells. The enrichments were plotted over the transcription start sites (TSS ± 2 kb) of protein-coding genes. Rows are sorted by decreasing H3K4me3 ChIP–seq occupancy in the control (DMSO RA 0h) cells.



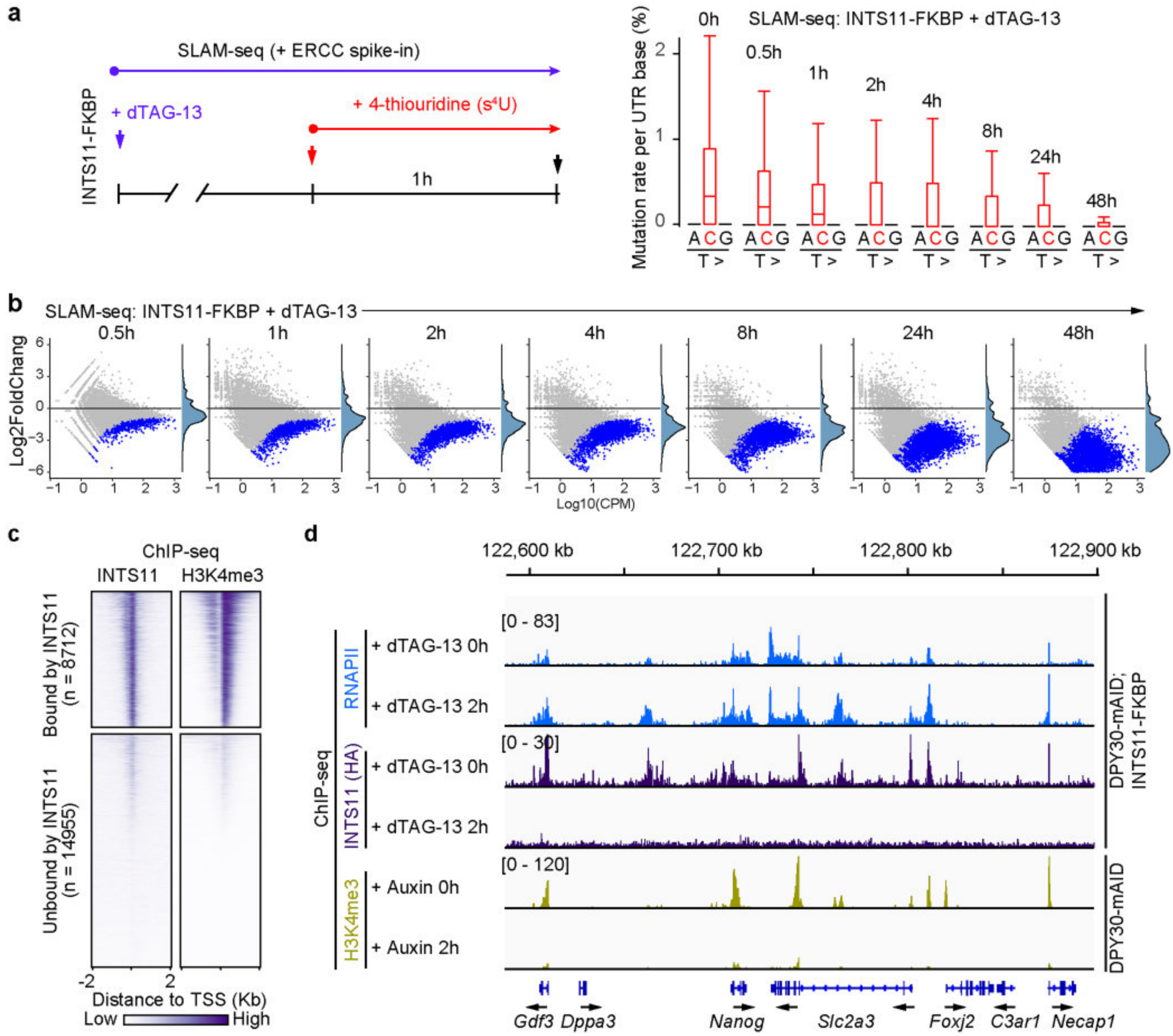
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | APEX2-based proteomic mapping scheme and characterization of endogenous RNAPII *in vivo* interactions. (**a**) Genomic confirmation of the APEX-modified *Rpb1* loci in DPY30–mAID and RBBP5–FKBP degron cells. (**b**) Sanger sequencing of the wild type and Flag-APEX2-RPB1 knock-ins. (**c**) Western blot showing the expression of APEX-modified RPB1 in DPY30–mAID and RBBP5–FKBP degron cells. (**d**) Representative brightfield images of mES cell colonies. (**e**) RT-qPCR analysis of the expression of pluripotency and differentiation genes in the knock-in cells. Data are from three biological replicates (n = 3) and are analysed using Two-way ANOVA and represented as mean ± s.d. (**f**) Titration of biotin phenol (BP). Cells stably expressing Flag-APEX2-RPB1 were pre-incubated for 30 min with the indicated concentrations of BP, followed by the addition of 1 mM H2O2 for 1 min. Cell lysates were probed with Streptavidin-HRP. Proximity biotinylation is optimal at a BP concentration of 4 mM. (**g**) Confirmation of the APEX2 functionality by protein biotinylation in the APEX2-engineered DPY30–mAID and RBBP5–FKBP degron cells. The discrete bands, denoted with asterisks, show APEX2-independent biotinylation by native enzymes. The Connexin-APEX2 overexpressed cell was severed as positive control for the APEX2 system. (**h**) SILAC-based chromatin proteomic strategy for mapping the neighbourhood interaction networks of APEX2-tagged RNAPII. (**i**) Principal component analysis (PCA) of SILAC signal in the RNAPII-APEX2 cells with or without agonist. (**j**) Distribution of SILAC ratio of RNAPII interactions quantified in the chromatin proteomic analyses. Mean log2 SILAC ratio is shown. In total, 1,901 proteins were identified in this experiment. The RNAPII-APEX2-bait (BP+H2O2) population has a right-shifted distribution compared with the no agonist negative control population, which indicates that the log2(SILAC) ratio allows us to distinguish bona fide RNAPII interactions from non-RNAPII interactions. (**k**) GO network showing significantly (q value < 0.001) enriched terms for positive RNAPII interaction neighbourhoods from RNAPII-APEX2 experiment. The most prominent pathways are indicated. Connecting lines show interaction of protein nodes. (**l**) KEGG enrichment and GO network showing significantly (q value < 0.001) enriched terms for positive RNAPII interaction neighbourhoods from RNAPII-APEX2 experiment. (**m**) Principal component analysis (PCA) of SILAC signal in the RNAPII-APEX2 DPY30–mAID cells with or without Auxin treatment. Time trajectory is shown by the dashed arrow. (**n**) Venn diagram indicating overlap between up or downregulated targets in the indicated samples. (**o**) Heatmap representing relative protein abundance of DPY30 and selected targets in Auxin treated DPY30–mAID cells. n = 3 independently samples. (**p**) Scatterplot analysis of proteins identified by SILAC in RNAPII-APEX2 DPY30–mAID cells following Auxin treatment for 2 and 8 h. (**q**) Gene ontology-based functional classification of 228 downregulated proteins in RNAPII-APEX2 DPY30–mAID cells following Auxin treatment for 2 and 8 h. The dot size is proportional to the number of members in an enrichment set, and colour intensity reflects the p value. Significance based on clusterProfiler analysis with Benjamini-Hochberg-adjusted P values.

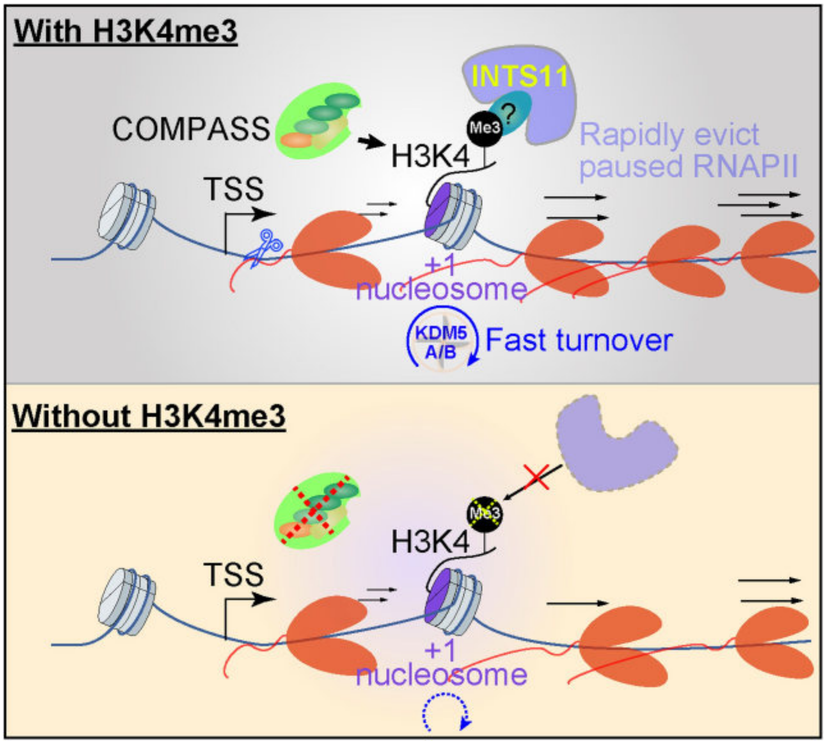


Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | INTS11 is required for nascent transcription. (**a**) Venn diagram indicating overlap of H3K4me3 interactors and RNAPII-APEX2 dependent interactors from ChIP-MS (chromatin proteomic profiling) data. (**b**) Relative enrichments of selected targets in various ChIP preparations based on ChIP-MS. (**c**) Validation of INTS11 interaction with H3K4me3 in RBBP5–FKBP degron cells at different times after dTAG-13 addition. Biotinylated proteins within lysates were enriched using Streptavidin-coated magnetic beads and analysed by Western blot. In parallel, sample in which H_2O_2 was omitted was prepared as negative control. (**d**) Schematic representation of the dTAG INTS11 targeting strategy for the INTS11–FKBP degron mES cells. (**e**) Western blot showing the expression of INTS11 and INTS11–FKBP–HA, using antibodies recognizing INTS11 or the HA in parental and knock-in degron cells. The arrow indicates the specific HA-tagged INTS11–FKBP–HA protein. (**f**) RT-qPCR analysis showing the expression of selected pluripotency and differentiation genes in the parental and INTS11–FKBP knock-in cells. Data are from three biological replicates (n = 3) and are analysed using Two-way ANOVA and represented as mean ± s.d. (**g**) Growth curve analysis of parental and INTS11–FKBP E14 cells treated with or without dTAG-13. (**h**) INTS11 enrichment profiles and heat maps as determined by using the HA-tag in control (0 h) and Auxin-treated (2 h) DPY30–mAID; INTS11–FKBP degron cells. Genome-wide binding averages showed enrichments at the TSS regions (TSS ± 2 kb) of protein coding genes. TSS, transcription start site. Rows were sorted by decreasing ChIP–seq occupancy in the control (0 h) cells. (**i**) Correlations between TT_{chem}-seq replicate experiments in INTS11–FKBP degron cells treated with or without dTAG-13 for the indicated times. (**j**) Average profiles for TT_{chem}-seq for the upstream anti-sense RNAs of each annotated protein-coding gene in INTS11 degron cells. TSS, transcription start site. (**k**) RNAPII profiles of various subclasses of annotations in INTS11–FKBP degron cells with or without dTAG-13 treatment. TSS, transcription start site. mRNA, messenger RNA. snRNA, small nuclear RNA. ncRNA, non-coding RNA. eRNA, enhancer RNA.



Extended Data Fig. 10 | Loss of INTS11 causes reduced transcriptional output of protein-coding genes. (**a**) Experimental design of SLAM-seq for INTS11–FKBP degron cells. Conversion rates for each position of a 4-thioU-containing SLAM-seq reads (≥ 2 T>C conversions) before or after dTAG-13 treatment. Average conversion rates (centre line) ± s.d. (whiskers) are shown. n = 3 biological replicates. The boxplot indicates the median (middle line) and the third and first quartiles (box); the whiskers show the 1.5× IQR above and below the box. (**b**) MA plots depicting changes in nascent transcription (SLAM-seq) at the indicated times after dTAG-13 treatment in INTS11–FKBP cells. n = 3 biological replicates. CPM, counts per million mapped reads. (**c**) H3K4me3 and INTS11 occupancy in mES cells. The enrichments were plotted over the transcription start sites (TSS ± 2 kb) of protein-coding genes. Rows are sorted by decreasing H3K4me3 ChIP–seq occupancy in the WT mES cells. (**d**) An IGV snapshot comparing ChIP–seq signals in control and dTAG-13-treated INTS11–FKBP cells.



Extended Data Fig. 11 | A Model for the roles of H3K4me3 in transcription regulation. H3K4me3 facilitates the recruitment of factors regulating the release of paused RNAPII at the +1 nucleosome. The H3K4me3 at promoter regions is highly dynamic, and it is maintained by an equilibrium between SET1/COMPASS complexes and KDM5 demethylases at highly transcribed genes. The rapid turnover of H3K4me3 ensures that the pausing step is a highly regulated process by Integrator Complex Subunit 11 (INTS11), where an increase in H3K4me3 leads to a decrease in RNAPII pausing and acute depletion leads to an increase RNAPII pausing.

nature portfolio

Corresponding author(s): KRISTIAN HELIN

Last updated by author(s): Dec 22, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Real Time quantitative PR data was collected on QuantStudio 6 Flex Real-time PR System v1.3 (Applied Biosystems). Next-generation sequencing data were collected via Illumina sequencing platforms (NextSeq 550). NextSeq550 control software (2.2.0) were used for high-throughput sequencing data collection.
Data analysis	The following software was used for data analysis: GraphPad Prism (v8.0.1) for general statistical analysis and graphing; FlowJo (v10.7.1) for flow cytometry analysis; Enrichr (v.2.1) for gene ontology (GO) enrichment; bcl2fastq (v2.19.0.316), FastQC (v0.11.8), Bowtie2 (v2.4.1), SAMtools (v1.10) , DeepTools (v3.3.0), HISAT (v2.2.1), SlamDunk (v0.3.4 16), DESeq2 (v1.34.0), cutadapt (v3.5) and STAR (v2.7.9a) for genomics data analysis; Proteome Discoverer (v3.1), MSstatsTMT (v2.4.0) and MaxQuant (v1.5.3.30) for proteomics data analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All related raw sequencing and related processed data are deposited and available from the Gene Expression Omnibus (GEO) under the accession numbers GSE181714.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No specific statistical measure was taken to decide sample size. Minimum sample sizes were predetermined from power estimates based on pilot experiments.
Data exclusions	No data has been excluded from analysis.
Replication	All attempts at replication were successful. Figure legends state how many times each experiment was performed.
Randomization	We did not carry out any randomization because this is either irrelevant or not applicable to this study.
Blinding	Blinding was not done since this study relies on the investigator studying differences in cell lines.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☐	☒	Antibodies
☐	☒	Eukaryotic cell lines
☒	☐	Palaeontology and archaeology
☒	☐	Animals and other organisms
☒	☐	Human research participants
☒	☐	Clinical data
☒	☐	Dual use research of concern

Methods

n/a	Involved in the study	
☐	☒	ChIP-seq
☐	☒	Flow cytometry
☒	☐	MRI-based neuroimaging

Antibodies

Antibodies used

anti-FLAG M2 Affinity gel SIGMA #A2220 50μl for each IP
anti-mini-AID-tag MBL #M2140-3 1:500 for WB
beta-ACTIN Abcam #ab6276 1:10000 for WB
BPTF Abcam #ab72036 1:1000 for WB
BRD4 Abcam #ab128874 1:1000 for WB; 2μg for ChIP
CDK7 Santa Cruz #sc-7344 1:500 for WB
CDK9 Santa Cruz #sc-13130 1:500 for WB; 2μg for ChIP
CHD1 Santa Cruz #sc-271626 1:500 for WB
CHD4 Abcam #ab240640 1:1000 for WB
DPY30 Bethyl Laboratories #A304-296A 1:1000 for WB; 2μg for ChIP
H3 Abcam #ab1791 1:10000 for WB

H3K36me3 Active Motif #61022 1:2000 for WB; 1µg for ChIP
H3K4me1 Cell Signalling #5326S 1:2000 for WB; 1µg for ChIP
H3K4me2 Cell Signalling #9725S 1:2000 for WB; 1µg for ChIP
H3K4me3 Cell Signalling #9751S 1:2000 for WB; 1µg for ChIP
HA-tag (C29F4) Cell Signaling #3724S 1:1000 for WB; 2µg for ChIP
HEXIM1 Cell Signaling #12604S 1:1000 for WB; 2µg for ChIP
INTS11 Atlas Antibodies #HPA029025 1:2000 for WB
KDM5A Cell Signaling #3876T 1:1000 for WB; 2µg for ChIP
KDM5B Helin Lab. Home made #Dain78 1:5000 for WB; 2µg for ChIP
NELFA Santa Cruz #sc-365004 1:500 for WB; 2µg for ChIP
NELFE Santa Cruz #sc-377052 1:500 for WB
PAF1 Abcam #ab137519 1:1000 for WB
Pol II Santa Cruz #sc-899; Abcam #ab817; Cell Signaling #2629S 1:1000 for WB; 2µg for ChIP
Pol II (phospho S2) Active motif #61984 1:1000 for WB; 2µg for ChIP
Pol II (phospho S5) Abcam #ab5131 1:1000 for WB; 2µg for ChIP
RBBP5 Bethyl Laboratories #A300-109A 1:1000 for WB; 2µg for ChIP
SPIN1 Abcam #ab118784 1:1000 for WB
SPT16 Santa Cruz #sc-377028 1:500 for WB
SPT6 Cell Signaling #15616S 1:1000 for WB
SSRP1 Cell Signaling #13421 1:1000 for WB
TAF3 Abcam #ab188332; EMD Millipore #07-1802 1:1000 for WB
INTS11 Atlas Antibodies AB #HPA029025 1:1000 for WB
TFIID (TBP) Santa Cruz #sc-421 1:500 for WB
FITC-conjugated goat α-Rabbit IgG antibody Invitrogen #F-2765 1:500 for Flow

Validation

Antibodies were validated by manufacturers or validated in previous studies. Statements on antibody validation are present on the manufacturer's websites along with relevant references. Additional validation was done by the use of negative control (control IgG) and control cells for Flow Cytometry analysis.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

E14 mESCs: DPY30-mAID:OsTiR1, DPY30-mAID:OsTiR1_Kdm5dKO, DPY30-mAID:OsTiR1_H3.3dKO, RBBP5-FKBP, DPY30-mAID:OsTiR1_RPB1-APEX2, RBBP5-FKBP_RPB1-APEX2

Authentication

Cells were not authenticated

Mycoplasma contamination

Cell lines were tested negative for mycoplasma

Commonly misidentified lines
(See ICLAC register)

No commonly misidentified cell lines were used

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

All related raw sequencing and related processed data are deposited and available from the Gene Expression Omnibus (GEO) under the accession numbers GSE181714.

Files in database submission

GSE181686 mNET-seq on COMPASS-degron cells
GSE181708 Quant-seq on COMPASS-degron cells
GSE181712 TT-seq on COMPASS-degron cells
GSE181892 ChIP-seq on COMPASS-degron cells

Genome browser session
(e.g. UCSC)

https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE181714&format=file

Methodology

Replicates

ChIP-seq was performed as individual replicates of two biologically independent degron systerms (Auxin and dTAG) mESCs. Information on reproducibility of the technique and antibody performance as described in the general section on replication of methods. Two replicates for mNET-seq and TT-seq.

Sequencing depth

The sequencing was performed in a NextSeq550. The sequencing read information are available as part of the GEO submission (GSE181714).

Antibodies

Target Source/Cat No. Application
BRD4 Abcam #ab128874 1:1000 for WB; 2µg for ChIP
CDK9 Santa Cruz #sc-13130 1:500 for WB; 2µg for ChIP

DPY30 Bethyl Laboratories #A304-296A 1:1000 for WB; 2µg for ChIP
H3K36me3 Active Motif #61022 1:1000 for WB; 1µg for ChIP
H3K4me1 Cell Signalling #5326S 1:1000 for WB; 1µg for ChIP
H3K4me2 Cell Signalling #9725S 1:1000 for WB; 1µg for ChIP
H3K4me3 Cell Signalling #9751S 1:1000 for WB; 1µg for ChIP
HEXIM1 Cell Signaling #12604S 1:1000 for WB; 2µg for ChIP
KDM5A Cell Signaling #3876T 1:1000 for WB; 2µg for ChIP
KDM5B Helin Lab. Home made #Dain78 1:5000 for WB; 2µg for ChIP
NELFA Santa Cruz #sc-365004 1:500 for WB; 2µg for ChIP
Pol II Santa Cruz #sc-899; Abcam #ab817; Cell Signaling #2629S 1:1000 for WB; 2µg for ChIP
Pol II (phospho S2) Active motif #61984 1:1000 for WB; 2µg for ChIP
Pol II (phospho S5) Abcam #ab5131 1:1000 for WB; 2µg for ChIP
RBBP5 Bethyl Laboratories #A300-109A 1:1000 for WB; 2µg for ChIP

Peak calling parameters: macs2 with Default parameters

Data quality: Data quality was assessed using fastqc

Software:
bcl2fastq (v2.19.0.316)
FastQC (v0.11.8)
Bowtie2 (v2.4.1)
SAMtools (v1.10)
DeepTools (v3.3.0)

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation: mESCs (1 million) were dissociated with Trypsin/EDTA, resuspended in culture medium, spun, and resuspended in PBS. 0.5 mL of cold Fixation Buffer (BioLegend, 420801) was added and then incubated at room temperature for 10 minutes. Subsequently, the cells were labeled with the unconjugated Rabbit DPY30 antibody (Bethyl Laboratories, A304-296A) and subsequently with a FITC-conjugated goat α-Rabbit IgG antibody.

Instrument: Beckman Coulter CytoFlex.

Software: FlowJo (v10.7.1)

Cell population abundance: Over 20,000 cells were counted for each sample

Gating strategy: The cells were gated using forward and side scatter parameters (FSC/SSC) for singlets (FSC-A/SSC-A). FL1-A :: B2-510-GFP-A is linked to DPY30 expression.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Article

Architecture of chloroplast TOC–TIC translocon supercomplex

https://doi.org/10.1038/s41586-023-05744-y

Received: 6 June 2022

Accepted: 19 January 2023

Published online: 26 January 2023

Check for updates

Hao Liu[1,2], Anjie Li[1,2], Jean-David Rochaix[3,4] & Zhenfeng Liu[1,2]✉

Chloroplasts rely on the translocon complexes in the outer and inner envelope membranes (the TOC and TIC complexes, respectively) to import thousands of different nuclear-encoded proteins from the cytosol[1–4]. Although previous studies indicated that the TOC and TIC complexes may assemble into larger supercomplexes[5–7], the overall architectures of the TOC–TIC supercomplexes and the mechanism of preprotein translocation are unclear. Here we report the cryo-electron microscopy structure of the TOC–TIC supercomplex from *Chlamydomonas reinhardtii*. The major subunits of the TOC complex (Toc75, Toc90 and Toc34) and TIC complex (Tic214, Tic20, Tic100 and Tic56), three chloroplast translocon-associated proteins (Ctap3, Ctap4 and Ctap5) and three newly identified small inner-membrane proteins (Simp1–3) have been located in the supercomplex. As the largest protein, Tic214 traverses the inner membrane, the intermembrane space and the outer membrane, connecting the TOC complex with the TIC proteins. An inositol hexaphosphate molecule is located at the Tic214–Toc90 interface and stabilizes their assembly. Four lipid molecules are located within or above an inner-membrane funnel formed by Tic214, Tic20, Simp1 and Ctap5. Multiple potential pathways found in the TOC–TIC supercomplex may support translocation of different substrate preproteins into chloroplasts.

As the organelles involved in photosynthesis, chloroplasts contain 2,000–3,000 nuclear-encoded proteins imported from the cytosol through the TOC and TIC complexes[8]. In the past decades, many different components of the TOC and TIC machineries have been identified using biochemical and genetic approaches[4,8]. The TOC complex mainly comprises three core subunits named Toc159, Toc34 and Toc75, whereas the TIC complex consists of four major subunits named Tic20, Tic214 (also known as Ycf1), Tic100 and Tic56 (ref. [9]). Whereas Toc159 and Toc34 function as the primary receptors for preproteins[2], Toc75 forms a translocation channel of the outer envelope by itself[10] or along with Toc159 and Toc34[11]. Tic20 serves as a central component of the TIC translocon[12], and Tic214 is essential for the import of various chloroplast proteins involved in photosynthesis, amino acid synthesis and ribosome biogenesis[7]. Tic100 and Tic56 are important for the assembly of a TIC translocon complex, as their absence led to a substantial reduction in the remaining TIC components[5]. Several other proteins, such as Tic236, Tic110, Tic62, Tic55, Tic40, Tic32, Tic22 and Tic21, have been reported as components of the TIC complex[2–4].

The TIC complex assembles with the TOC complex to form a large translocon supercomplex with a molecular mass of over 1 MDa (refs. [5–7,13]). The TOC–TIC supercomplexes have been found in both land plants (*Arabidopsis thaliana* and *Pisum sativum*)[5,6] and a green alga (*C. reinhardtii*)[7]. Notably, the major components of the TIC (Tic20, Tic21, Tic32 and Tic110) and TOC (Toc75, Toc34 and Toc159) complexes are present in both green and red lineages[4,14], whereas Tic214 is conserved in most green-lineage species except for grasses, some parasitic plants and cranberry[15]. Although the protein-import function and components of the TOC–TIC supercomplex have been investigated extensively, little is known about the assembly mechanism of the different subunits of the supercomplex, and the translocation pathway for the preproteins remains unclear.

Overall architecture

The TOC–TIC supercomplex sample, purified from a *C. reinhardtii* strain expressing a Tic20–Venus–3×Flag fusion protein[16], was used for structural analysis using the single-particle cryo-electron microscopy (cryo-EM) method. As shown in Fig. 1a, the TOC–TIC supercomplex (at a resolution of 2.8 Å) from *C. reinhardtii* has an overall shape resembling a torch, and it is composed of three complexes, namely TOC, TIC and the intermembrane space complex (ISC) connecting TOC and TIC. TOC mainly contains Toc75, Toc34, Toc90 (a homologue of plant Toc159) and a small unidentified chain (chain X) embedded in the outer membrane. Flanking on the Toc90 side, a β-barrel protein named Ctap4 is located at a closest distance of around 12 Å from TOC (Fig. 1b). For TIC, the transmembrane domains of Tic214 and Tic20 form the core region surrounded by the transmembrane helices (TMHs) of Ctap5, Simp1, Simp2 and Simp3 (Fig. 1c). In the intermembrane space, Tic100, Tic56 and the soluble domains of Ctap3, Ctap5, Simp1 and Simp2 intertwine with the intermembrane space domains of Tic214 to form the ISC, providing an extended surface to accommodate the intermembrane space domains of Toc90, Toc75, Toc34 and Ctap4. The results of purification,

[1]National Laboratory of Biomacromolecules, CAS Centre for Excellence in Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China. [2]College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China. [3]Department of Molecular Biology, University of Geneva, Geneva, Switzerland. [4]Department of Plant Biology, University of Geneva, Geneva, Switzerland. ✉e-mail: liuzf@ibp.ac.cn

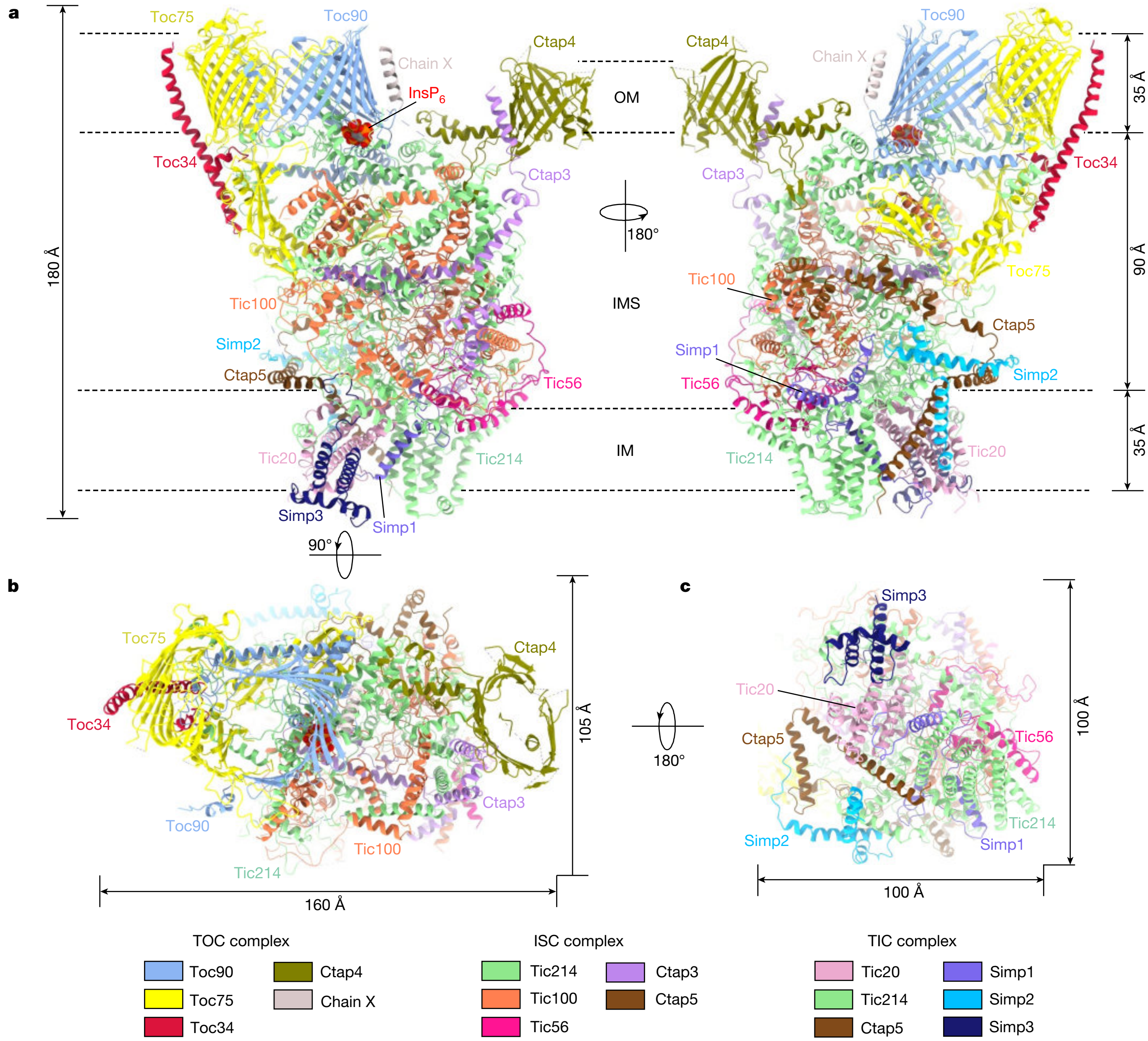


Fig. 1 | The overall architecture of the TOC–TIC supercomplex. a, Side views of the supercomplex model along the membrane planes. The protein subunits are shown as cartoon models. OM, the outer membrane of chloroplast envelope; IM, the inner membrane; IMS, the intermembrane space. **b**, Top view of the supercomplex from the cytosolic side. **c**, Bottom view of the supercomplex from the chloroplast stromal side. The $InsP_6$ molecule at the interface between Toc90 and Tic214 is highlighted as a sphere model.

characterization and protein identification of the supercomplex sample are summarized in Extended Data Fig. 1, Supplementary Data 1 and Supplementary Tables 1 and 2. Previously, incubation of *Chlamydomonas* chloroplasts with pre-ferredoxin (pre-Fdx1) and 0.3 mM ATP led to an affinity-purified translocation intermediate complex containing most of the proteins found in the structure except Simp3[7] (Supplementary Table 3). The cryo-EM data collection, processing and structure refinement results are presented in Extended Data Fig. 2 and Extended Data Table 1.

Notably, the three-dimensional (3D) variability and multibody analysis results indicate that the Toc75–Toc90–Toc34 complex and Ctap4 are mobile relative to the ISC and TIC regions (Supplementary Videos 1 and 2). For example, they may rotate slightly as rigid bodies by about 3° around a pivot point in the upper side of the ISC region. Ctap4 has a concerted movement along with the Toc75–Toc90–Toc34 complex, as an α-helical domain of Tic214 fills in their gap and binds the two outer-membrane units together (Supplementary Video 1). According to the results of the multibody analysis (Supplementary Video 2), the Toc75–Toc90–Toc34 complex and Ctap4 exhibit at least nine different motion components. They can either move together in a synchronized way or separately as independent bodies. The flexibility of the TOC–TIC supercomplex may enable it to adjust the gap between the TOC and TIC complexes so as to facilitate the translocation of preproteins across the intermembrane space.

The translocon in the outer membrane

Although previous studies suggested that Toc75 can form a channel by itself[10] or in a complex with Toc159 and Toc34 (ref. [11]), the cryo-EM structure shows that Toc75 assembles closely with Toc90 at a 1:1 ratio

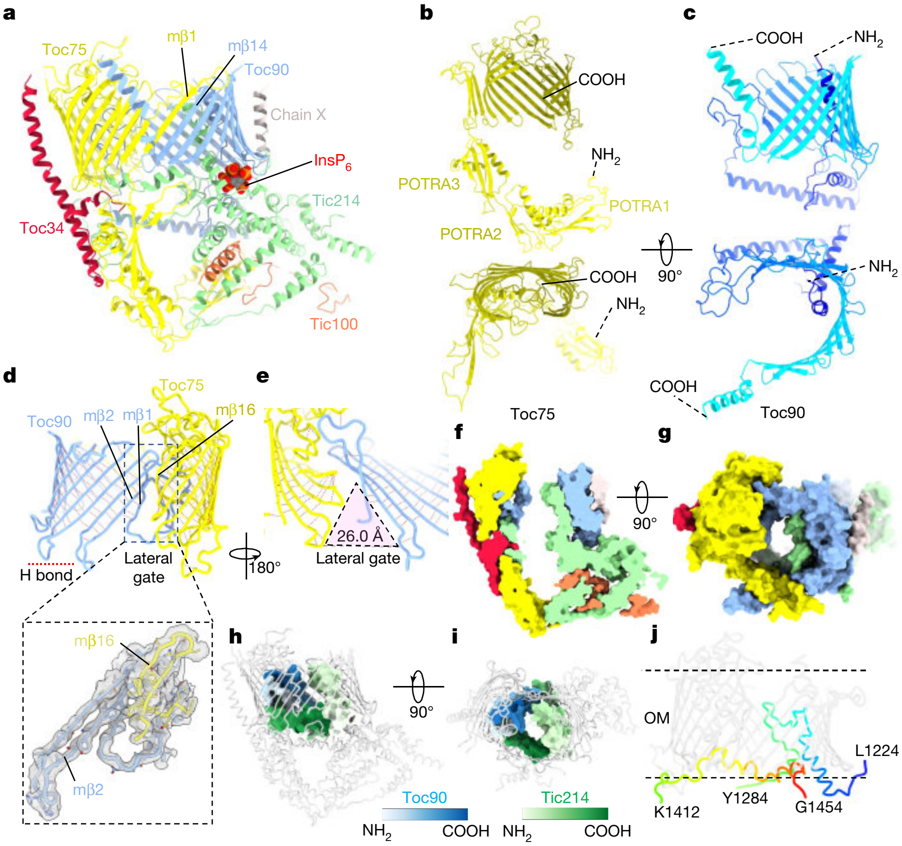


Fig. 2 | Structure of the Toc75–Toc90–Toc34 complex associated with Tic214. **a**, Cartoon model of the Toc75–Toc90–Toc34–Tic214 complex. **b**, Side view and top view of Toc75 structure. **c**, Side view and top view of Toc90 structure. **d**,**e**, A potential lateral gate located between mβ16 of Toc75 and mβ2 of Toc90. The views are from the external (**d**) and internal (**e**) sides of the β-barrel. The hydrogen bonds (H bonds) between the backbones of adjacent β-strands are indicated by the dotted lines. The pink dashed triangle labels the lower half of the lateral gate. **f**,**g**, Sectional side view (**f**) and top view (**g**) of the surface presentation of the Toc75–Toc90–Toc34–Tic214 complex. The colour code is the same as in **a**. **h**,**i**, The proteins inside the pore lumen are presented as surface models and coloured in gradient modes (Toc90, white to blue; Tic214, white to green). h, side view; i, top view. **j**, Two motifs of Tic214 located inside the TOC channel formed by Toc75 and Toc90. The dashed lines indicate the approximate locations of the membrane surfaces.

to form the channel in the outer membrane (Fig. 2a). Toc34 binds to the peripheral region of Toc75 instead of participating directly in channel formation. The amino-terminal region of Toc75 contains three polypeptide-transport-associated domains (POTRA1–3) which are followed by a membrane-embedded open β-barrel domain with 16 antiparallel β-strands (Fig. 2b). POTRA2 is tilted by approximately 58° relative to POTRA1, whereas POTRA2 forms a smaller angle (around 35°) with POTRA3. The relationship between POTRA1 and POTRA2 differs largely from the crystal structure of POTRA domains from *A. thaliana* Toc75 (ref. [17]), whereas the POTRA2–POTRA3 domains of the two structures superpose well (Extended Data Fig. 3). In the TOC–TIC supercomplex, POTRA1 is stabilized by nearby proteins (Toc90, Tic214 and Tic100) and forms specific interactions with them, whereas, in the crystal structure, it is surrounded by symmetry-related molecules. Such a difference may also suggest that the hinge between the POTRA1 and POTRA2 domains is flexible and could allow for domain rearrangements at different stages of preprotein translocation. The transmembrane domain (TMD) of Toc75 is curved (Fig. 2b) and assembles with the TMD of Toc90 to form an enclosed channel (Fig. 2a and Extended Data Fig. 4).

As one of the four members (Toc159, Toc132, Toc120 and Toc90) of the Toc159 family[4], plant Toc90 has a protein-import function similar to that of Toc159[18,19] and may prefer photosynthetic proteins, such as the light-harvesting proteins (LHCP)[20]. The TMD of *C. reinhardtii* Toc90 contains 14 antiparallel β-strands (Fig. 2c) and forms a C-shaped structure to enclose the adjacent one of Toc75. The β-barrel domain of Toc90 assembles specifically with Toc75 on two different sides, forming a heterodimeric TOC channel within the outer membrane. On one side, the last membrane-embedded β-strand (mβ14) of Toc90 interacts closely with the first membrane-embedded β-strand (mβ1) of Toc75 through backbone hydrogen bonds, forming an antiparallel β-sheet (Fig. 2a). On the other side, the first and second membrane-embedded β-strand (mβ1 and mβ2) of Toc90 form weak van der Waals interactions with the last membrane-embedded β-strand (mβ16) of Toc75 (Fig. 2d,e). Such a weak contact site may potentially function as a flexible lateral gate, opening for transport of preproteins targeted to the outer membrane, or when the preprotein is too large to fit in the central pore. A recent study demonstrated that an extra-superfolded green fluorescent protein (esGFP) might be translocated through TOC–TIC translocons without unfolding[21]. As the size of folded GFP (about 30 Å wide and 40–50 Å tall) is much larger than the central pore of the TOC channel, a transient opening of the lateral gate may be necessary for the channel to accommodate esGFP and enable it to pass through.

The amino-proximal regions of Toc159 and Toc34 contain a GTP-binding domain (G domain) forming a heterodimer or homodimer in vitro and in vivo[22,23] (Extended Data Fig. 5a,b).

Notably, the Toc75–Toc90 channel contains a hollow pore in the middle (Fig. 2f,g), extending from the cytosolic side to the intermembrane space side. The pore lumen surface is mainly lined by amino acid residues from Toc75 (mβ1–mβ5) and Toc90 (α3–mβ1 loop, mβ1–mβ4, mβ13–mβ14) (Extended Data Fig. 4c,f). The α3–mβ1 loop and mβ1–mβ2 loop of Toc90 insert into the cavity of Toc75, and fill the pore lumen of the TOC complex on the Toc75 side (Fig. 2h,i (blue)). Moreover, it is

notable that Tic214 also participates in forming the pore lumen surface of the TOC complex by extending two motifs (Leu1224–Tyr1284 and Lys1412–Gly1454) inside the pore (Fig. 2h,i (green)). Whereas the first motif contains a π-shaped helix-loop structure extending from the intermembrane space to the cytoplasmic surface, the second motif contains two short α-helices located near the exit to the intermembrane space (Fig. 2j (rainbow)). They may contribute to substrate recognition and interaction during translocation of the preproteins through the channel.

Flanking on the Toc90 side, the small β-barrel protein is assigned as Ctap4 (Fig. 1a and Extended Data Fig. 5c–g). In addition to the β-barrel domain embedded in the outer membrane, the protein contains two hydrophilic motifs (HM1 and HM2) in the intermembrane space region. Previously, it was found that seven Ctap proteins (Ctap1–7), including Ctap4, were copurified along with the TOC–TIC supercomplex from *C. reinhardtii*[7]. Our results demonstrate that Ctap4 is a β-barrel membrane protein located in the outer membrane that interacts closely with Tic214 and Ctap3. An α-helix (Val268–Glu283) of Ctap3 is attached to the outer surface of the Ctap4 β-barrel, whereas the elongated C-terminal domain of Ctap3 is located in the intermembrane space region (Fig. 1a).

An $InsP_6$ molecule

In an electropositive pocket on the intermembrane-space side of the TOC complex, a small molecule is surrounded by positively charged residues and sandwiched between Trp1235 of Tic214 and Ser768 of Toc90 (Extended Data Fig. 6a–c). The cryo-EM density matches well with the model of an inositol hexaphosphate ($InsP_6$) molecule (Extended Data Fig. 6b). The $InsP_6$ molecule is from the endogenous source of *C. reinhardtii* cells and co-purified along with the TOC–TIC supercomplex. The polar extract of small molecules from the purified TOC–TIC supercomplex sample exhibits features that are similar to those of the $InsP_6$ standard sample in mass spectrometry (MS) analysis, confirming that the TOC–TIC supercomplex does contain $InsP_6$ (Extended Data Fig. 6f,g). The $InsP_6$ closely interacts with seven positively charged residues from Tic214 and Toc90 through salt bridges and four other residues through hydrogen bonds (Extended Data Fig. 6c–e). Such strong interactions secure the assembly of the pore-lining motif of Tic214 with the inner surface of Toc90 on the intermembrane-space side, as $InsP_6$ wedges in the space between Tic214 and Toc90.

Inositol polyphosphates are phosphorylated derivatives of myo-inositol and have crucial roles in the regulation of metabolic processes, phosphate storage and sensing, hormone signalling and regulation of photosynthetic function in higher plants and algae[24,25]. The discovery of $InsP_6$ in the TOC–TIC supercomplex suggests that $InsP_6$ is important for the assembly of the TOC–TIC supercomplex, as the binding site of $InsP_6$ is at a pivotal interfacial position in the supercomplex.

The ISC

In the space between the outer and inner membranes, Tic100, Tic214, Tic56, Ctap3 and Ctap5 assemble into a tower-shaped complex measuring around 80–90 Å in height (Fig. 3a–c). The upper half of the ISC facing the outer membrane forms a distorted triangular bowl with a shallow cavity in the middle and positive charges enriched at the mouth of the bowl (Fig. 3b), whereas the lower half facing the inner membrane forms a wider quadrilateral base structure (Fig. 3c). The ISC has a crucial role in connecting TOC with the inner-membrane domains of TIC (Extended Data Fig. 7a–e).

Among the five proteins involved in forming the ISC, Tic214 is the largest one that interacts with nearly all of the other proteins in the supercomplex. As shown in Fig. 3d, the Tic214 structure mainly consists of around 41 α-helices, two pairs of short antiparallel β-strands and numerous irregular loops. Whereas the amino-proximal region of Tic214 forms a compact TMD embedded in the inner membrane, the middle region folds into an extended and loose structure (intermembrane-space domain (ISD); Extended Data Fig. 7b) spanning 70–80 Å across the intermembrane space to connect the TMD with the carboxy-proximal motifs inside the pore lumen of the Toc75–Toc90 channel. The extended structure of Tic214 enables it to connect TOC with TIC proteins and provides a flexible platform for other proteins to bind. The gene encoding Tic214 is essential for the survival of *C. reinhardtii*[26] and higher plant[27] cells, and recent biochemical studies indicate that it is a constitutive component of the TOC–TIC supercomplexes[5,7]. Thus, Tic214 is not only the central subunit for ISC, but is also essential for the assembly of the TOC–TIC supercomplex in general.

Tic100 and Tic56 are both soluble proteins that are important for the assembly of the TIC complex and essential for the accumulation of photosynthetic proteins in plants[5]. Tic100 has an overall shape of a wreath with four major domains, namely the amino-terminal α-helical domain (NTAD), β-jellyroll domain (BJD), middle α-helical domain (MAHD) and the carboxy-terminal α-helical domain (CTAD) (Fig. 3e and Extended Data Fig. 7c). Instead of forming a compact structure by itself, Tic100 intertwines with the ISD of Tic214 to form the central part of the ISC. The *Arabidopsis tic100* mutant has a reduced level of 1 MDa TIC complex and a lower chloroplast protein import rate in comparison to the wild type, and mutation of a glycine residue (corresponding to Gly418 in *C. reinhardtii* Tic100, located near to a membrane occupation and recognition nexus motif in the BJD domain; Fig. 3e) to Arg led to a partial loss of function of Tic100[28]. The mutation may affect the assembly between Tic100 and Tic214 as Gly418 of Tic100 is located at their interface. Thus, Tic100 is crucial for the assembly of the TIC complex and may facilitate the process of protein translocation by establishing a physical bridge between the TOC and TIC complexes (Fig. 3f).

Tic56 is a smaller protein with a short α-helix (Gly78–Thr93, α1) at the N-terminal region connected with a carboxy-proximal globular domain (Pro127–Arg244) through a long loop (Extended Data Fig. 7d). Whereas the α1 of Tic56 interacts with the carboxy-terminal regions of Tic100 and Ctap3, the carboxy-proximal globular domain of Tic56 intercalates at the cleft between the TMD of Tic214 and the CTAD of Tic100 to stabilize their assembly (Fig. 3a). The absence of Tic56 in *Arabidopsis* also leads to a reduction in the 1 MDa TIC complex and the chloroplasts are deficient in protein import[5,29].

The translocon in the inner membrane

Tic20 is a crucial component of the TIC complex, interacting closely with a preprotein substrate during the late stage of the import process[30]. Mutant plants with reduced Tic20 levels exhibit inefficient import and accumulation of plastid preproteins, reduced photosynthetic capacity and growth defects[12,31]. As shown in Fig. 4a–c, Tic20 assembles with the TMD of Tic214 and four small membrane proteins (Ctap5, Simp1, Simp2 and Simp3) to form the translocon complex embedded in the inner membrane. There are 15 TMHs in TIC, arranged in an asymmetric manner within the membrane. Tic20 and Tic214 contain four and six TMHs, respectively (Fig. 4d,e), whereas the remaining five are from Ctap5, Simp1, Simp2 and Simp3 (Extended Data Fig. 8). At the peripheral region of Tic20, Simp3 associates with TMH3 and α4 of Tic20 to stabilize the local structure around Tic20 (Fig. 4b). Simp3 also interacts closely with Tic214. The carboxy-terminal tail of Simp3 attaches to a surface pocket surrounded by the three parts (Pro468–Ile475, Arg735–Lys749 and Lys815–Tyr825) of Tic214, and a short helix–loop–helix motif (Arg426–Lys447) of Tic214 connects Simp3 (TMH2) with Tic20 (TMH3–α3).

Notably, TMH1 and TMH4 from Tic20, TMH1 from Simp1, TMH3 and TMH6 from Tic214, and TMH1 from Ctap5 collectively form a funnel-like structure in the middle region, resembling the SecY-protein-conducting channel[32] (Fig. 4f,g). On the intermembrane-space side, the funnel is covered by a long loop region (Gly328–Val380) following TMH6 of Tic214 and the α1–α2a loop region of Tic20 (Fig. 4a). Instead of having

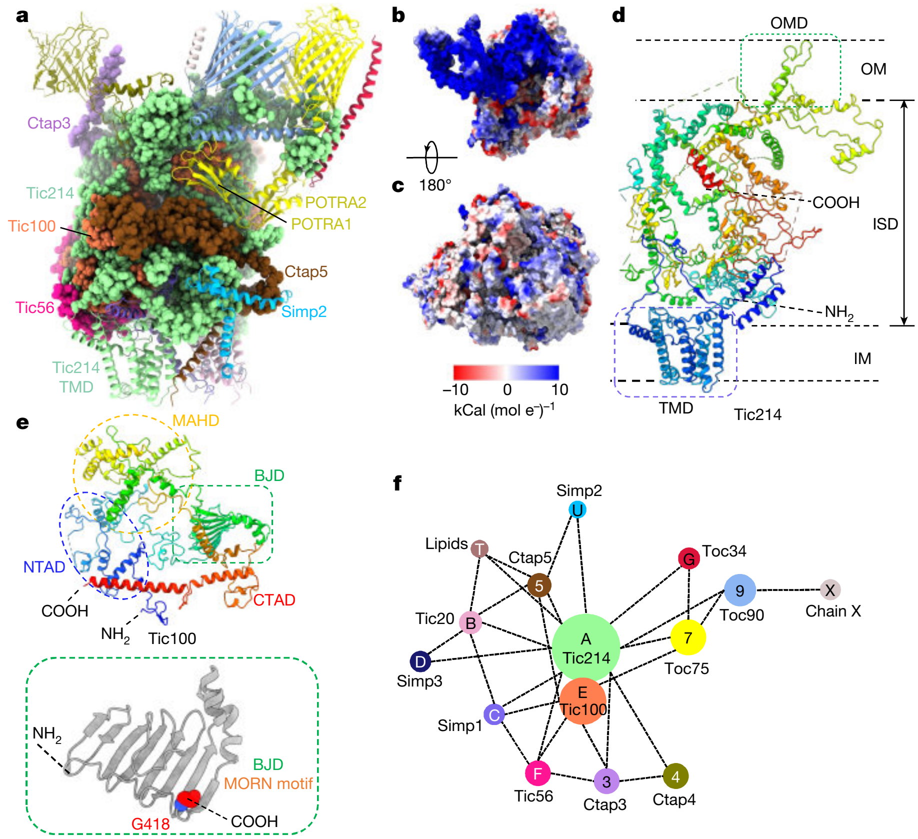


Fig. 3 | Structural role and components of the ISC. a, The ISC provides binding sites on the surface for the outer-membrane and inner-membrane proteins. The proteins involved in forming the ISC are shown as sphere models, and the outer-membrane and inner-membrane proteins are shown as cartoon models. **b**,**c**, Top view (**b**) and bottom view (**c**) of the ISC from the outer-membrane and inner-membrane sides, respectively. The proteins are presented as surface models, coloured by the level of coulombic electrostatic potential (blue, electropositive; red, electronegative; white, neutral). **d**,**e**, The overall structure of Tic214 (**d**) and Tic100 (**e**). The proteins are coloured in the rainbow mode. Blue, the amino-terminal region; cyan, green and yellow, middle region; red, the carboxy-terminal region. The red sphere in **e** highlights a glycine residue that is crucial for the in vivo function of Tic100 as reported previously. OMD, outer-membrane domain of Tic214; ISD, ISD of Tic214; TMD, TMD of Tic214 embedded in the inner membrane. MORN motif, the membrane occupation and recognition nexus motif found in Tic100 and potentially involved in binding to lipids from the membrane. **f**, The central role of the Tic214–Tic100 complex in forming close interactions with the other proteins of the supercomplex. Each protein is indicated as a circle and the single-letter codes on the circles are the chain codes for the structural models of the corresponding proteins. The interactions between adjacent proteins are indicated by the dotted lines.

a substrate protein inside the funnel, three lipid molecules are located inside the pore and a fourth one is found in a pocket above the funnel (Fig. 4h and Extended Data Fig. 9a). The lipid molecules form hydrogen bonds and hydrophobic interactions with amino acid residues from Tic20, Ctap5, Tic214 and Simp1 (Extended Data Fig. 9b,c). At the bottom side facing the stroma, a lipid molecule (PG604) serves as a plug to occlude the bottleneck and prevent leakage of the funnel (Fig. 4h). Moreover, the funnel has a lateral gate opened towards the lipid bilayer (Fig. 4f). A similar lateral gate was also found in SecY, functioning as an exit for the hydrophobic signal peptide segment and/or the membrane-spanning domains of substrate proteins[32] (Fig. 4g). The lateral gate in the TIC complex is located between TMH3 of Tic214 and TMH1 of Ctap5, whereas the one in SecY lies between TMH2 and TMH7 from the same chain. While the two pairs of gate-lining helices intertwine in different ways (right-handed for the TIC channel versus left-handed for SecY), they both form V-shaped lateral gates connected with lipid bilayers. The lateral gate of SecY provides a hydrophobic surface groove for binding of a hydrophobic α-helix of the signal sequence[32]. Similarly, the lateral gate of the TIC channel also forms a surface groove that is mainly lined by hydrophobic residues (such as Leu159 from Ctap5, Phe186 from Tic214), potentially functioning to bind to the hydrophobic segments of some transit peptides or other regions of preproteins transiently during the translocation process. Moreover, TMH1 of Simp2 is located at the edge of the lateral gate, defining the peripheral boundary of the exit gate (Fig. 4b). Moreover, numerous other lipid (and detergent) molecules are located at the peripheral region of the TIC complex (Extended Data Fig. 9d).

Putative protein translocation pathways

TOC contains a straight open pore running from the cytosolic side to the intermembrane space (Fig. 5a). The pore appears to have a bottleneck near the cytoplasmic entrance, measuring 11.7 Å wide in the shortest dimension and 20.5 Å wide in the longest dimension (Fig. 5b). It is surrounded by amino acid residues from the mβ1–mβ2 loop region of Toc90 as well as the mβ3–mβ4 loop and mβ7–mβ8 of Toc75 (Fig. 5a). The surface of the pore lumen is electropositive on one side and electronegative on the other side (Extended Data Fig. 10a,b). While the entrance appears to be wide enough to accommodate an extended polypeptide of the substrate preprotein, the flexible loops around the entrance may adjust dynamically according to different parts of the translocated preprotein. Below the entrance, the pore diameter increases to 14–22 Å, wider than the entrance (Fig. 5b). There are two exit portals on different sides of the bottom (Fig. 5c,d (exits 1 and 2)). Whereas exit 1 is open directly to the intermembrane space, exit 2 is connected to the exit 3 route in a slide-like groove formed by POTRA2–POTRA3 of Toc75 and

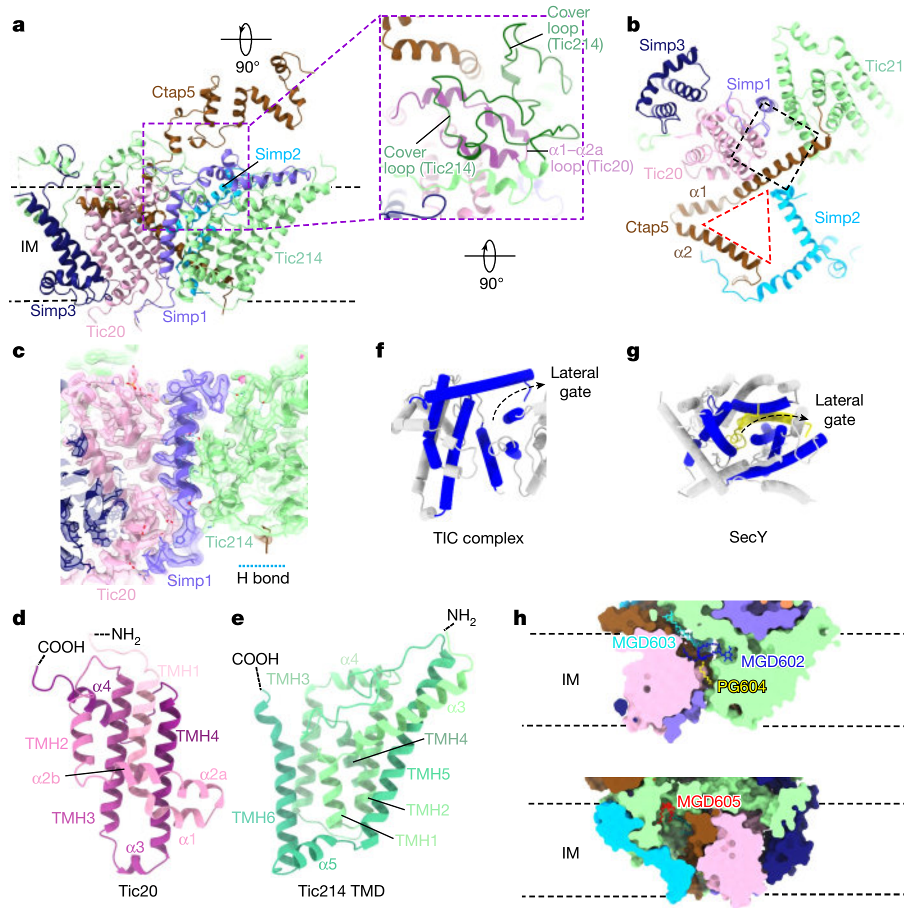


Fig. 4 | Arrangement of protein subunits and lipid molecules in the inner-membrane complex. a, Side view (**a**) and bottom view (**b**) of the translocon complex in the inner membrane. For **a**, the area in the purple dashed box contains the loops from Tic20 and Tic214 covering the funnel entrance underneath. The view in **b** is from the stromal side and the area in the dark dashed box contains a central funnel filled with lipid molecules (omitted for clarity). The red dashed triangle indicates the entrance formed by the α1 and α2 of Ctap5 on the intermembrane space side. **c**, The interface between Tic20 and Tic214. The structural models are superposed with corresponding cryo-EM densities shown in transparent mode. **d**,**e**, The structures of Tic20 (**d**) and the Tic214 TMD (**e**). **f**, The central funnel-like region of the Tic20–Tic214–Ctap5–Simp1 complex. The view is from the intermembrane space side. **g**, The protein-conducting channel of SecY with substrate bound. The TMHs encircling the central pore are coloured in blue, and the other parts are shown in silver. The polypeptide substrate of SecY is shown in yellow (Protein Data Bank (PDB): 5EUL). **h**, Binding sites of four lipid molecules at the subunit interfaces. MGD, monogalactosyl-diacylglycerol; PG, phosphatidyl glycerol.

the nearby parts of Tic100, Tic214 and Ctap3 (Fig. 5c,e and Extended Data Fig. 10c,d). Such a groove may help to guide the preproteins to the cavity above the central funnel of the TIC complex or to the adjacent portal through the triangular entrance by Ctap5.

The pore in the central funnel between Tic214 and Tic20 is overall much narrower than the one in the TOC complex (Fig. 5f), indicating that the TIC channel may be in a closed state. The upper half of the pore is 7.0–12.1 Å wide, whereas the lower half is only 5.2–6.6 Å wide (Fig. 5g). The constriction site is located near to the exit at the stromal surface and surrounded by amino acid residues from Tic214, Ctap5 and Tic20 (Fig. 5f). Thus, the TIC pore will need to expand by adjusting local protein conformations substantially for the preprotein to enter and go through. As the TMHs of Ctap5 and Simp1 are not tightly restrained by nearby subunits, they might be able to move outwards and adjust flexibly in response to the preprotein being translocated.

The TOC and TIC complexes are responsible for translocation of various nucleus-encoded proteins into different chloroplast subcompartments[1,3]. As different substrate preproteins have distinct surface properties and target locations, they might be translocated through different pathways. For soluble proteins targeted to the stroma, such as the small subunit of ribulose 1,5-bisphosphate carboxylase or ferredoxin[13,30], they will need to pass through the central pore of the Toc75–Toc90 channel first. Either guided by the surface groove on ISC or facilitated by Tic236 (with a large ISD)[33], they may efficiently traverse the intermembrane space and are further translocated across the inner membrane, presumably through the central funnel within the Tic20–Tic214–Ctap5–Simp1 complex (Fig. 5e (exit 3-1)). As the central funnel of the Tic20–Tic214–Ctap5–Simp1 complex is covered by the flexible loops from Tic214 and Tic20 (Fig. 4a), opening of the entrance will probably require rearrangement of the cover loop for the preproteins to enter. The molecular dynamics simulations of the TIC complex embedded in a lipid bilayer suggest that the cover loop from Tic214 is fairly flexible and may be involved in binding to the transit peptide region of the preprotein, whereas its spontaneous movement is not large enough to open the entrance (Supplementary Videos 3 and 4). The central pore may still need to expand to accommodate the translocated preproteins, so that they can pass through the pore and reach the stroma. Alternatively, the preproteins might bypass the central funnel and use a side portal connected with the central funnel (Fig. 5e (exit 3-2)), after going through a triangle-shaped entrance outlined by two amphipathic helices (α1 and α2) of Ctap5 (Fig. 4b and Extended Data Fig. 10d,e). Such a side portal along the surface of TMH1 of Ctap5 might support translocation of photosynthetic membrane proteins (such as LHCP) and others, by providing transitional docking sites for their transmembrane/hydrophobic

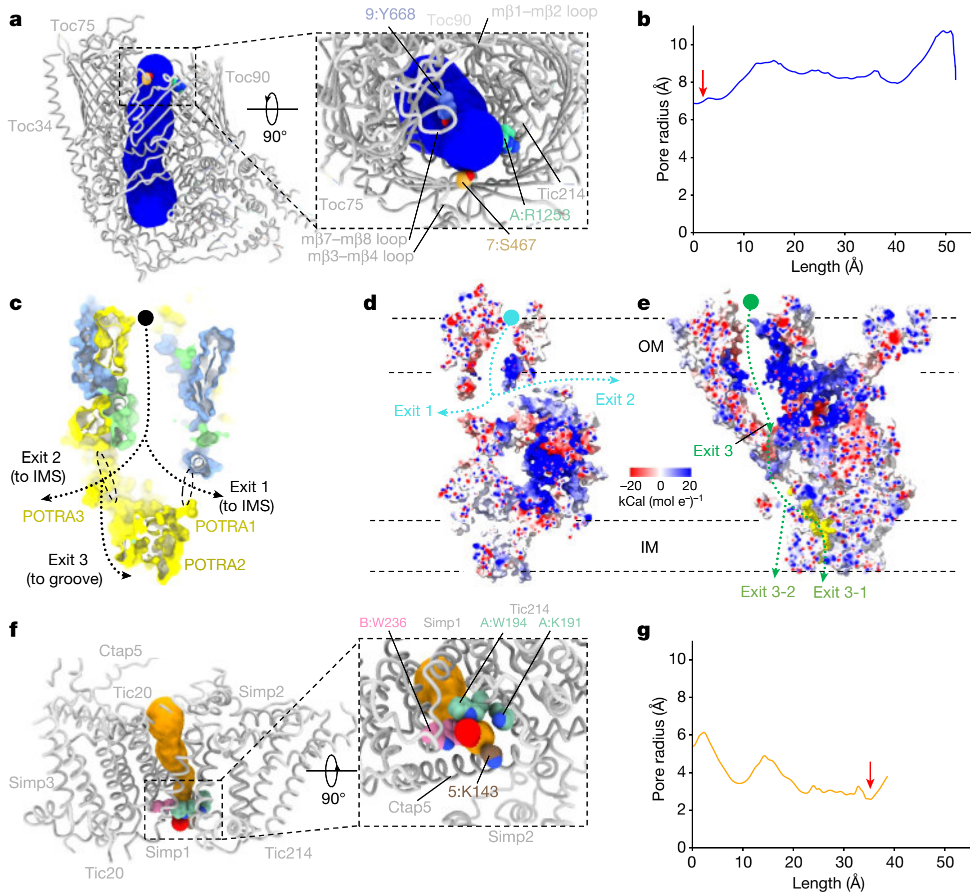


Fig. 5 | The intrinsic pores found in the TOC and TIC complexes. a, The presence of a continuous wide pore running through the TOC complex. The putative pore is presented as a transparent blue tube model. **b**, The distribution of the pore radius along the central axis of the TOC pore. The red arrow indicates the constriction site of the pore with the smallest pore radius. **c**, The presence of two exit portals at the bottom of the TOC complex. The two exit portals are indicated by the dashed elliptical rings. Note that exit 2 and exit 3 share the same portal at the bottom of TOC complex on the opposite side of exit 1 portal. **d**,**e**, The potential translocation pathways for proteins targeted to the intermembrane space (**d**) and the soluble proteins or LHCP proteins targeted to stroma (**e**). The cyan and green dotted lines indicate the approximate locations of the translocation pathways. The TOC and TIC proteins are presented as surface models coloured by the level of Coulombic electrostatic potential (blue, electropositive; white, neutral; red electronegative), and lipid molecules are shown as yellow spheres. **f**, The central pore of the TIC complex shown as a transparent orange tube model. The region in red is the constriction site. The key amino acid residues at the constriction site near to the stromal surface (**f**) or at the cytoplasmic entrance of TOC (**a**) are presented in sphere models and labelled by chain name:residue name and number. A, Tic214; B, Tic20; 5, Ctap5; 7, Toc75; 9, Toc90. **g**, The distribution of pore radius along the central axis of the TIC pore. The red arrow indicates the constriction site of the pore with the smallest pore radius.

domains. For the preproteins targeted to the intermembrane space, such as the Tic22 protein[34], they may exit the Toc75–Toc90 channel through a side portal (exit 1) outlined by the POTRA1–POTRA2 of Toc75 and α3 of Toc90, or through the adjacent one (exit 2) between POTRA2–POTRA3 of Toc75 and the bowl-like structure of the Tic214–Tic100 complex (Fig. 5d). Tic236 might also participate in the import of Tic22 preprotein into chloroplasts according to a recent functional study[35].

Discussion

The cryo-EM map reveals a TOC complex from *C. reinhardtii* composed of Toc90, Toc75 and Toc34 with a stoichiometry of 1:1:1, whereas previous studies on plant TOC complexes suggested that the stoichiometry of Toc159:Toc75:Toc34 is 1:4:4–5 or 1:3:3 (ref. [11,36]). Thus, much larger TOC complexes may exist in plants. Alternatively, the Toc159–Toc75–Toc34 complexes with a 1:1:1 ratio might coexist with the Toc75–Toc34 complexes with no Toc159 bound, accounting for the excess molar ratio of Toc75 and Toc34 relative to Toc159. It is also possible that additional Toc75 and Toc34 subunits are associated with the TOC–TIC supercomplex and were lost during the sample preparation process for the cryo-EM study.

Previously, it was reported that Toc75 or Tic20 alone can form channels by itself when reconstituted in liposomes[37,38], whereas our research demonstrates that Toc75 assembles with Toc90 to form the heteromeric TOC channel and is further connected with the TIC complex. Nevertheless, it cannot be ruled out that Toc75 or Tic20 may still form channels alone, in case of free Toc75 or Tic20 proteins in the membrane.

Although the cryo-EM structure of TOC–TIC supercomplex from *C. reinhardtii* includes three Toc proteins, four Tic proteins, three Ctap proteins and three Simp proteins, several other TIC components identified previously are still absent in the structure (Supplementary Table 3). It was reported that Tic236 functions as a large protein linking the TOC and TIC complexes[33], whereas no density corresponding to Tic236 can be found in the map. Notably, Tic214 spans both envelope membranes and, together with Tic100, instead of Tic236, forms the ISC core to link the TOC and TIC complexes (Fig. 3a,f). Tic214 is the only chloroplast-encoded protein of the TOC–TIC supercomplex and may function as a scaffolding protein.

Previous research suggested that Tic110 may form a channel that is sensitive to transit peptides and function as a translocation pore for the preproteins at the inner membrane[39]. Another biochemical study on plant TIC complex indicated that Tic21 is loosely associated with

the complex[13]. In the cryo-EM map of the TOC–TIC supercomplex from *C. reinhardtii*, neither Tic110 nor Tic21 is observed, suggesting that they are either lost during purification or may exist in very low abundance. Alternatively, Tic110 may function downstream of the translocon channel instead of being part of the TOC–TIC supercomplex[8].

When a tightly folded protein is used as the substrate for the TOC–TIC translocon, the pore may open to a size greater than 25.6 Å (ref. [40]), much larger than the pores observed in the structure (Fig. 5a,f). In that case, the TOC and TIC complexes will need to undergo large conformational changes in the pore-lining regions to accommodate the folded proteins. Recently, a Ycf2–FtsHi heteromeric AAA-ATPase complex was identified to be the import motor associated with the TIC complex from *A. thaliana*, using the energy from ATP to drive the import process[41]. In *C. reinhardtii*, three FtsH-like AAA-proteins and an orthologue of Ycf2 named Orf2971 associated with Tic20 and Tic214 were also identified[7,42]. In the TIC complex, there is a large concaved groove between the TMD of Tic214 and Simp3 (Extended Data Fig. 10f). Potentially, the groove and the U-shaped stromal surface of the TIC complex may provide binding sites for the import motor. In the cryo-EM map, no extra densities corresponding to the FtsH-like AAA proteins are visible, suggesting that they might also be lost during the purification process or their association with the TIC complex relies on the presence of preprotein substrates. Thus, the TOC–TIC supercomplex structure reported here represents the pre-import state without any preprotein substrates or AAA-ATPase proteins bound.

A recent functional study demonstrated that Tic12, a plant homologue of Simp1 and Simp2, is an essential component of the TIC complex, and both Tic12 and Tic20 interact with the transit peptide of a translocating preprotein[43]. Consistently, our structure reveals that Simp1 flanks the central funnel of the Tic20–Tic214–Ctap5 complex on one side, whereas Simp2 functions as a peripheral fence on the other side nearby Ctap5 (Fig. 4b). While our work was under review, a study reported the structure of *Chlamydomonas* TOC–TIC supercomplex, and proposed a preprotein translocation pathway of TIC complex located between Tic20 and YlmG (or Simp3)[44]. This pathway is on a peripheral surface groove of the TIC complex and distant from the central funnel of Tic20–Tic214–Ctap5–Simp1 (named Tic12 in ref. [44]) complex. Notably, there is lack of functional or biochemical evidence supporting the involvement of YlmG in forming the preprotein translocation pathway, whereas the central funnel-like pathway that we propose is consistent with the previous functional studies[5,7,30,43]. The two exit portals of TOC open to the intermembrane space (Fig. 5d) were not described in ref. [44], Toc90 was named as Toc120 instead and most of the transmembrane helix (TMH1) of Simp2 (or Tic13 in ref. [44]) was not detected (see Supplementary Table 3 for details on the nomenclature differences). There are two potential *Toc159* orthologues in the *Chlamydomonas* nuclear genome, namely *Cre17.g734300* and *Cre17.g707500*. *Cre17.g734300* encodes a protein with 967 amino acid residues and a molecular mass of 102,303 Da, smaller than the protein (1,080 amino acid residues; molecular mass, 108,935 Da) encoded by *Cre17.g707500*. The closest homologue of Cre17.g734300 found in the Non-Redundant Protein Sequence database of *A. thaliana* (through the BLASTp search) is Toc90. *At*Toc90 shares 28% identities and 44% positives (query coverage, 63%; score, 172; *E* value, 1×10^{-43}) with the query sequence of Cre17.g734300. By contrast, *A. thaliana* Toc120 versus Cre17.g734300 alignment exhibits much lower query coverage (34%) and score (168) as well as a higher *E* value (8×10^{-42}). Thus, the Cre17.g734300 protein is more similar to *At*Toc90 than to *At*Toc120 and, to be consistent with the previous work[7], it is named Toc90 instead of Toc120.

In our cryo-EM map of the TOC–TIC supercomplex, there are some extra densities attached to the side chain hydroxyl groups of specific serine or threonine residues in Tic100 (Ser173, Ser229, Thr239, Ser244, Thr249, Ser251, Thr510, Ser514, Thr523, Ser706, Thr708, Thr712 and Ser714), Toc90 (Thr564), Ctap3 (Thr328) and Tic214 (Thr1795). These densities may potentially belong to the phosphoryl groups of phosphorylated Thr/Ser residues as reported by ref. [44]. Alternatively, they may also belong to water molecules or ions. The potential phosphorylation sites in Tic100, Toc90, Ctap3 and Tic214 will be verified through further biochemical and functional studies.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-05744-y.

1. Soll, J. & Schleiff, E. Protein import into chloroplasts. *Nat. Rev. Mol. Cell Biol.* **5**, 198–208 (2004).
2. Richardson, L. G. L. & Schnell, D. J. Origins, function, and regulation of the TOC-TIC general protein import machinery of plastids. *J. Exp. Bot.* **71**, 1226–1238 (2020).
3. Li, H. M. & Chiu, C. C. Protein transport into chloroplasts. *Annu. Rev. Plant Biol.* **61**, 157–180 (2010).
4. Shi, L. X. & Theg, S. M. The chloroplast protein import system: from algae to trees. *Biochim. Biophys. Acta* **1833**, 314–331 (2013).
5. Kikuchi, S. et al. Uncovering the protein translocon at the chloroplast inner envelope membrane. *Science* **339**, 571–574 (2013).
6. Chen, L. J. & Li, H. M. Stable megadalton TOC-TIC supercomplexes as major mediators of protein import into chloroplasts. *Plant J.* **92**, 178–188 (2017).
7. Ramundo, S. et al. Coexpressed subunits of dual genetic origin define a conserved supercomplex mediating essential protein import into chloroplasts. *Proc. Natl Acad. Sci. USA* **117**, 32739–32749 (2020).
8. Rochaix, J.-D. Chloroplast protein import machinery and quality control. *FEBS J.* **289**, 6908–6918 (2022).
9. Nakai, M. New perspectives on chloroplast protein import. *Plant Cell Physiol.* **59**, 1111–1119 (2018).
10. Hinnah, S. C., Hill, K., Wagner, R., Schlicher, T. & Soll, J. Reconstitution of a chloroplast protein import channel. *EMBO J.* **16**, 7351–7360 (1997).
11. Schleiff, E., Soll, J., Küchler, M., Kühlbrandt, W. & Harrer, R. Characterization of the translocon of the outer envelope of chloroplasts. *J. Cell Biol.* **160**, 541–551 (2003).
12. Chen, X., Smith, M. D., Fitzpatrick, L. & Schnell, D. J. In vivo analysis of the role of atTic20 in protein import into chloroplasts. *Plant Cell* **14**, 641–654 (2002).
13. Kikuchi, S. et al. A 1-megadalton translocation complex containing Tic20 and Tic21 mediates chloroplast protein import at the inner envelope membrane. *Plant Cell* **21**, 1781–1797 (2009).
14. Kalanon, M. & McFadden, G. I. The chloroplast protein translocation complexes of *Chlamydomonas reinhardtii*: a bioinformatic comparison of Toc and Tic components in plants, green algae and red algae. *Genetics* **179**, 95–112 (2008).
15. de Vries, J., Sousa, F. L., Bölter, B., Soll, J. & Gould, S. B. YCF1: a green TIC? *Plant Cell* **27**, 1827–1833 (2015).
16. Mackinder, L. C. M. et al. A spatial interactome reveals the protein organization of the algal CO_2-concentrating mechanism. *Cell* **171**, 133–147 (2017).
17. O'Neil, P. K. et al. The POTRA domains of Toc75 exhibit chaperone-like function to facilitate import into chloroplasts. *Proc. Natl Acad. Sci. USA* **114**, E4868–E4876 (2017).
18. Infanger, S. et al. The chloroplast import receptor Toc90 partially restores the accumulation of Toc159 client proteins in the *Arabidopsis thaliana* ppi2 mutant. *Mol. Plant* **4**, 252–263 (2011).
19. Bauer, J. et al. The major protein import receptor of plastids is essential for chloroplast biogenesis. *Nature* **403**, 203–207 (2000).
20. Hiltbrunner, A. et al. AtToc90, a new GTP-binding component of the *Arabidopsis* chloroplast protein import machinery. *Plant Mol. Biol.* **54**, 427–440 (2004).
21. Jeong, J., Moon, B., Hwang, I. & Lee, D. W. Green fluorescent protein variants with enhanced folding are more efficiently imported into chloroplasts. *Plant Physiol.* **190**, 238–249 (2022).
22. Wiesemann, K., Simm, S., Mirus, O., Ladig, R. & Schleiff, E. Regulation of two GTPases Toc159 and Toc34 in the translocon of the outer envelope of chloroplasts. *Biochim. Biophys. Acta* **1867**, 627–636 (2019).
23. Rahim, G., Bischof, S., Kessler, F. & Agne, B. In vivo interaction between atToc33 and atToc159 GTP-binding domains demonstrated in a plant split-ubiquitin system. *J. Exp. Bot.* **60**, 257–267 (2008).
24. Couso, I. et al. Inositol polyphosphates and target of rapamycin kinase signalling govern photosystem II protein phosphorylation and photosynthetic function under light stress in *Chlamydomonas*. *New Phytol.* **232**, 2011–2025 (2021).
25. Lorenzo-Orts, L., Couto, D. & Hothorn, M. Identity and functions of inorganic and inositol polyphosphates in plants. *New Phytol.* **225**, 637–652 (2020).
26. Boudreau, E. et al. A large open reading frame (orf1995) in the chloroplast DNA of *Chlamydomonas reinhardtii* encodes an essential protein. *Mol. Gen. Genet.* **253**, 649–653 (1997).
27. Drescher, A., Ruf, S., Calsa, T. Jr, Carrer, H. & Bock, R. The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. *Plant J.* **22**, 97–104 (2000).
28. Loudya, N. et al. Mutations in the chloroplast inner envelope protein TIC100 impair and repair chloroplast protein import and impact retrograde signaling. *Plant Cell* **34**, 3028–3046 (2022).
29. Köhler, D. et al. Characterization of chloroplast protein import without Tic56, a component of the 1-megadalton translocon at the inner envelope membrane of chloroplasts. *Plant Physiol.* **167**, 972–990 (2015).

30. Kouranov, A. & Schnell, D. J. Analysis of the interactions of preproteins with the import machinery over the course of protein import into chloroplasts. *J. Cell Biol.* **139**, 1677–1685 (1997).
31. Hirabayashi, Y., Kikuchi, S., Oishi, M. & Nakai, M. In vivo studies on the roles of two closely related *Arabidopsis* Tic20 proteins, AtTic20-I and AtTic20-IV. *Plant Cell Physiol.* **52**, 469–478 (2011).
32. Li, L. et al. Crystal structure of a substrate-engaged SecY protein-translocation channel. *Nature* **531**, 395–399 (2016).
33. Chen, Y.-L. et al. TIC236 links the outer and inner membrane translocons of the chloroplast. *Nature* **564**, 125–129 (2018).
34. Vojta, L., Soll, J. & Bölter, B. Protein transport in chloroplasts—targeting to the intermembrane space. *FEBS J.* **274**, 5043–5054 (2007).
35. Chuang, M.-R., Chen, L.-J. & Li, H.-M. Chloroplast import of an intermembrane space protein is facilitated by translocon components Toc75 and Tic236. *Plant Direct* **5**, e356 (2021).
36. Kikuchi, S., Hirohashi, T. & Nakai, M. Characterization of the preprotein translocon at the outer envelope membrane of chloroplasts by blue native PAGE. *Plant Cell Physiol.* **47**, 363–371 (2006).
37. Kovács-Bogdán, E., Benz, J. P., Soll, J. & Bölter, B. Tic20 forms a channel independent of Tic110 in chloroplasts. *BMC Plant Biol.* **11**, 133 (2011).
38. Hinnah, S. C., Wagner, R., Sveshnikova, N., Harrer, R. & Soll, J. The chloroplast protein import channel Toc75: pore properties and interaction with transit peptides. *Biophys. J.* **83**, 899–911 (2002).
39. Heins, L. et al. The preprotein conducting channel at the inner envelope membrane of plastids. *EMBO J.* **21**, 2616–2625 (2002).
40. Ganesan, I., Shi, L.-X., Labs, M. & Theg, S. M. Evaluating the functional pore size of chloroplast TOC and TIC protein translocons: import of folded proteins. *Plant Cell* **30**, 2161–2173 (2018).
41. Kikuchi, S. et al. A Ycf2-FtsHi heteromeric AAA-ATPase complex is required for chloroplast protein import. *Plant Cell* **30**, 2677–2703 (2018).
42. Xing, J. et al. The plastid-encoded protein Orf2971 is required for protein translocation and chloroplast quality control. *Plant Cell* **34**, 3383–3399 (2022).
43. Zhao, X., Higa, T. & Nakai, M. Tic12, a 12-kDa essential component of the translocon at the inner envelope membrane of chloroplasts in *Arabidopsis*. *Plant Cell* **34**, 4569–4582 (2022).
44. Jin, Z. et al. Structure of a TOC-TIC supercomplex spanning two chloroplast envelope membranes. *Cell* **185**, 4788–4800 (2022).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2023

Methods

Purification and biochemical analysis

The *C. reinhardtii* CS1_FC1D12 strain made in ref. [16] from the *Chlamydomonas* Resource Center expressing Tic20–Venus–3×Flag fusion protein (the target gene was cloned into pLM005 vector and transformed into a wild-type strain cMJ030/CC-4533) was used for purification of the TOC–TIC supercomplex using affinity[7] and size-exclusion chromatography (SEC) methods. The Venus–3×Flag is fused to the carboxy-terminal region of Tic20 and does not interfere with the assembly of the TOC–TIC supercomplex, as the Flag-tag affinity-purified sample contains components similar to the functional translocon intermediates active in importing preprotein substrates[7].

For protein expression, the single colony selected with 25 μg ml^{-1} paromomycin was used to inoculate 100 ml Tris-acetate-phosphate medium with 4× Kropat's Trace Elements[45]. After the cell density reached an optical density at 600 nm (OD_{600}) of 1.0, the culture was then used to inoculate 2 l TAP medium with constant bubbling of filtered air through pumps at 27 °C under 150 μmol m^{-2} s^{-1} irradiance for 6–7 days. The cells in stationary phase with an OD_{600} of 1.0–1.3 were collected by centrifugation at 5,053*g* in the JLA-8.1 rotor (Beckman) and the pellets were stored at −80 °C.

Throughout the protein-purification process, all of the steps were performed at 4 °C. For each preparation, 50 g frozen cell pellets were resuspended in a lysis buffer (50 mM HEPES-KOH pH 7.5, 250 mM NaCl, 10% (v/v) glycerol, 5 mM EDTA, 4 mM dithiothreitol (DTT), 1× protease inhibitor cocktail (EDTA-free, 100× in DMSO, MedChemExpress/MCE)) at 1:10 (m/v) ratio, yielding a suspension with about 1 mg ml^{-1} chlorophyll. The suspension was homogenized by using the T10 basic homogenizer (IKA). The cells were lysed by passing through a high-pressure homogenizer (ATS Engineering, Shanghai) 5 times at 1,000 bar. Lauryl maltose neopentyl glycol (LMNG, NG310, Anatrace) and digitonin (D-3203, BIOSYNTH) were added to the cell lysate at 1.0% (w/v) and 0.2% (w/v) final concentration, respectively, to solubilize the membrane protein complexes. 1 ml MonoRab Anti-DYKDDDDK Magnetic Beads (L00835, GenScipt) were added to the mixture to capture the target protein. After the mixture was stirred for 2 h, the magnetic beads were isolated and washed at least five times using the wash buffer (50 mM HEPES-KOH pH 7.5, 150 mM NaCl, 10% (v/v) glycerol, 1 mM DTT, 0.06% digitonin). The target protein complexes were eluted twice with 500 μg ml^{-1} 3× DYKDDDDK peptide (RP21087, GenScript) dissolved in the wash buffer. The eluent was concentrated to about 0.5 ml and then applied to the SEC column (Superose 6, 10/300, GE Healthcare) in the SEC buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 0.06% Digitonin). The fractions containing the TOC–TIC supercomplex were pooled and concentrated to 16 μl in a 100 kDa MWCO Amicon Ultra-0.5 Centrifugal Filter Unit (UFC5100, Millipore).

Protein identification

For protein identification, the TOC–TIC supercomplex sample was analysed firstly through sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS–PAGE, using ExpressPlus PAGE Gel, 10×8, 4–12%, GenScipt), or through the blue native-polyacrylamide gel electrophoresis (BN-PAGE)[46]. The gels were fixed and stained using the Silver Stain for Mass Spectrometry kit (C500021, Sangon Biotech). Each individual protein band from the SDS–PAGE or BN-PAGE was excised and silver was removed before the MS analysis.

For preparing the peptide sample from the protein bands of SDS–PAGE, the protein samples were treated with DTT and iodoacetamide, and trypsin was added to the samples to digest the proteins overnight. Subsequently, the products were desalted by using C18 ZipTip column, mixed with Alpha-cyano-4-hydroxycinnamic acid (CHCA) and then spotted on the plate. The data were acquired through the matrix-assisted laser desorption/ionization time-of-flight/time-of-flight mass spectrometer (MALDI-TOF/TOF UltraFlexTreme). Database search was performed by using the Peptide Mass Fingerprint or MS/MS ION Search page from a local MASCOT website. Cysteine carbamidomethylation was set as a fixed modification, while methionine oxidation was set as a variable modification. The SWISSPROT database was used to establish a reference database for MALDI-TOF MS spectra.

For peptide samples prepared from the BN-PAGE bands, the samples were analysed using the nano LC-Q EXACTIVE system (Thermo Fisher Scientific). The protein samples were treated with DTT and iodoacetamide, and then trypsin enzymatic solution was added. The mixtures were incubated for 30 min, and then chymotrypsin was added to digest the proteins overnight. The peptides were next separated by passing the samples through the C18 reversed-phase column (7.5 μm × 20 cm, 3 μm) at a flow rate of 300 nl min^{-1}. The mobile phase consisted of two components, namely phase A (0.1% formic acid aqueous) and phase B (0.1% formic acid acetonitrile). The column was first eluted with a linear gradient from 4% to 8% B from 0 to 8 min and then with a linear gradient from 22% to 32% B from 8 min to 58 min, followed by an equilibration step for 7 min with 90% B. The MS analysis was performed with a scan range of 300–1,600 *m*/*z* for MS, ion spray voltage was set at 2.0 kV in the positive mode, and the micropipette temperature was 320 °C. The data collected were used for searching against the UniProt_Proteome_Chlamydomonas reinhardtii_2021 database to find matching fragments using Thermo Proteome Discoverer (v.1.4.0.288).

Grid preparation and data collection

For cryo-EM sample preparation, 3.0 μl of the supercomplex sample was added onto a H_2/O_2 glow-discharged (in Solarus 950, Gatan) holey carbon grid coated with 2 nm ultrathin carbon film (Quantifoil 300-mesh, R 2/2). After waiting for 60 s, the grids were blotted for 8 s with a force level of 0 and immersed into liquid ethane using a Vitrobot (Mark IV, Thermo Fisher Scientific) under 100% humidity at 4 °C. The frozen grids were then transferred to liquid nitrogen for storage.

The images were collected using a system comprising the Titan Krios G3 (Thermo Fisher Scientific) operating at 300 kV and a K2 Summit detector (Gatan). Video stacks were automatically acquired in the super-resolution mode (with super-resolution pixel size 0.675 Å) using the SerialEM program, with a defocus range from −1.5 μm to −1.8 μm. The total dose was approximately 50 e^{-} $Å^{-2}$ in 33 frames. The images were recorded through the beam-image shift data collection method[47].

Cryo-EM data processing and refinement

The workflow for cryo-EM data processing is summarized in Extended Data Fig. 2. The procedures described below were performed using cryoSPARC (v.3.2)[48] unless stated otherwise. First, 9,984 video stacks were motion-corrected by Patch-based motion correction and dose-weighted with twofold binning. Micrograph contrast transfer functions were estimated by Patch-based CTF estimation.

For particle picking, an initial round was performed using manual picking and the Blob Picker Tuner. After the first round of 2D classification, ab initio reconstruction and non-uniform refinement[49] were carried out. The 3D map exhibited a severe preferred orientation due to lack of top-view and bottom-view particles. To pick the particles with more orientations, simulated data were used to generate particles with all orientations. Subsequently, a 2D classification step using the simulated data was used to generate templates for the Template picking process. Around 6,814,089 particles were picked using template picking and extracted with fourfold binning.

For 3D classification in Relion (v.4.0)[50], all particles were converted to star format using cs2star (v.0.4.1). 3D classification (classes = 6, T = 6) was carried out to generate six reference maps using the refined model obtained in the last non-uniform refinement step as a reference. All of the maps were imported to cryoSPARC, and the corresponding selected particles were re-extracted with twofold binning. After a heterogeneous refinement and a non-uniform refinement job, a reconstruction at an overall resolution of 2.9 Å was obtained. Then, two rounds of

heterogeneous refinement followed by local CTF refinement, global CTF refinement, remove duplicate particles, local motion correction and non-uniform refinement were performed and further improved the overall map resolution to 2.77 Å.

As the region around the TOC complex has much weaker densities compared with the TIC complex region, a mask covering the TOC complex and nearby outer membrane domains was generated using UCSF ChimeraX (v.1.4)[51] and Volume Tools in cryoSPARC, and applied to improve the local map quality. A combined process with local refinement and global CTF refine was carried out in two rounds to generate a local map at 3.40 Å resolution. To further improve the local map quality, a combination of particle subtraction and local refinement steps was carried out and a local map around the TOC complex (Toc34, Toc75 and Toc90) and Ctap4 at 3.03 Å resolution was obtained (Extended Data Fig. 2e). The mask used for particle subtraction covers the bulk regions of ISC and TIC complexes except for the TOC complex, Ctap4 and the closely associated parts of Tic214. Following the particle subtraction process, the output data were subject to local refinement with the mask covering the entire TOC complex and Ctap4. After the refinement, the map quality around the TOC complex in the local map appears much better than the corresponding region in the global map.

The cryo-EM maps were sharpened using Sharpen tools in cryoSPARC and with the *B* factor reported in non-uniform refinement or local refinement. All of the map figures were generated using ChimeraX. To analyse the dynamic features of the individual components in the TOC–TIC supercomplex, the non-uniform-refined particles were binned at 2.5, and imported into Relion (v.4.0) using cs2star.py. Auto-refine was first performed in Relion (v.4.0), leading to a density map with a resolution of 3.6 Å. Three rigid bodies were inferred from the 3D variability analysis result output by cryoSPARC (Supplementary Video 1), namely the TOC complex (Toc34, Toc75 and Toc90), the ISC–TIC complex and Ctap4. Tight masks covering the three bodies were created and used for the multibody refinement and analysis. The initial sampling angle was 1.8°, the initial offset was 3 px and the step size is 0.75 px. The motion of the three rigid bodies is decomposed along 18 different eigenvectors. The first 9 eigenvectors explain 89.7% of the variance in the data and are significantly higher than the other eigenvectors. They were exported, aligned on the ISC–TIC part and used for generating animation videos in ChimeraX (Supplementary Video 2).

Model building, refinement and analysis

For model building, the initial model of each individual protein subunit of the TOC–TIC supercomplex was predicted using AlphaFold2 (ref. [52]). The predicted models were used as references or docked directly in the cryo-EM map (if they have features fitting well with the densities) during the model building process. The sharpened maps were used as references to guide the model building process, while the original maps were used for model refinement. All identifiable parts of the predicted models matching with the local density features were selected and placed into the map by fitting them as rigid bodies in the map using UCSF Chimera (v.1.16)[53] or WinCoot (v.0.9.4)[54]. Subsequently, the models were manually extended, adjusted and rebuilt in WinCoot by referring to the amino acid sequences, secondary structure prediction as well as the local densities. For Tic20 and Tic214, the predicted models of their TMDs were docked first into the corresponding region of the cryo-EM map and adjusted manually to fit each amino acid residue in the map. Afterwards, the models were extended at the N-terminal and C-terminal regions according to the amino acid sequences and the map features. For Tic214, the large ISD has a structure largely different from the predicted model. To build its model, characteristic motifs of different regions, such as the QWWRTKQR motif in the 1233–1240 region and the RQWWW motif in the 1475–1479 region, were identified in the map. Their models were built ab initio and then extended at the two termini according to the sequence and map features. For Tic100, the characteristic BJD was identified first in the map and the predicted model of this part was fitted into the map, adjusted and extended at both ends. For Tic56, Ctap3 and Ctap5, segments of their predicted models in the middle regions were identified in the map first and then docked in the map, adjusted and extended. The three small inner membrane proteins (Simp1, Simp2 and Simp3) were identified by referring to the predicted models of small proteins identified in the MS analysis and manual fitting of the characteristic segments of the predicted models in the map. Once a good fit was achieved, the model was completed by extending on both ends according to the sequence and continuous local map features.

For the models of TOC proteins, Toc34 was identified by its characteristic long transmembrane helix and the model was extended on both ends. The predicted model of Toc75 was adapted into the map by adjusting the three POTRA domains individually and the β-strands in the membrane-spanning domain. Toc90 was identified by two α-helices in the intermembrane space domain. The predicted model of this part was fitted in the map, and then combined with the adapted model of the β-barrel domain, which was adjusted and fitted strand-by-strand in the map. Ctap4 was identified by three characteristic α-helices in the intermembrane space region (Leu223–Leu259) and a β-barrel domain in the outer-membrane region. After the Toc34, Toc75, Toc90, Ctap4 and Chain X (unidentified due to insufficient local map resolution) were individually built in the local map, they were combined and merged with the TIC complex model. All the other non-specific proteins detected using MS (Supplementary Table 2) were also examined one by one, and they do not exhibit structural features matching with the densities in the cryo-EM map.

As for the lipid molecules found in the structure, the MGDG molecules (MGD602, MGD603 and MG605) are recognized by the characteristic ring-shaped head groups, whereas the PG molecules (PG604, PG2007 and so on) are modelled by considering their relatively small head-group densities and compatible local environments. Whereas the density for PG2007 head group is fairly strong and fits the PG model well, those for PG604 and others (tentatively assigned as PG, but may also belong to phosphatidylethanolamine or 1,2-diacylglyceryl-trimethylhomoserine) are relatively weak because they are located at the peripheral regions with high flexibility.

The supercomplex model was further adjusted and refined in WinCoot manually against the global map. For model refinement, electronic Ligand Builder and Optimization Workbench (eLBOW)[55] was used to generate restraint information for small-molecule ligands. The models were refined using phenix.real_space_refine[56]. The refined model was checked and readjusted in WinCoot manually. The model building and refinement cycle was repeated multiple times until most of the map features were interpreted by the model and the comprehensive validation statistical indicators for structural models were satisfied.

For channel pore analysis, the structural models of TOC or TIC complexes were imported to CAVER Analyst 2.0 (ref. [57]). The TOC complex model with Toc34, Toc75, Toc 90 and part of Tic214 was used for probing the pore inside the TOC channel. Moreover, the TIC complex model with Tic20, Tic214, Simp1, Simp3 and Ctap5 was used for the search of a potential pore inside the TIC channel. The channel pores were then selected, exported and converted to BILD format for visualization in ChimeraX. The electrostatic potential surface representations were calculated by using ChimeraX.

Extraction and identification of $InsP_6$

For extraction of $InsP_6$ (also known as phytic acid) molecules from the purified TOC–TIC supercomplex sample, 140 mg ml^{-1} biobeads (SM2, Bio-Rad) were added to the sample, mixed and incubated overnight at 25 °C. The mixture was centrifuged at 17,000*g* for 10 min at 25 °C. The supernatant was removed by pipetting, and 50% (v/v) glycerol was added to the pellets so that biobeads could be separated from the precipitated proteins. Subsequently, the mixture was centrifuged again, and the precipitate was retained. The precipitate was diluted by

adding water to the same volume as the supernatant. NaOH (10 M) and 500 mM EDTA Na were added to the suspension to final concentrations of 250 mM and 50 mM, respectively. After being dried under vacuum, the sample was dissolved in a mixture of methanol and water (1:1, v/v). The mixture was centrifuged at 17,000*g* for 1 min at 25 °C. The supernatant was collected immediately and used for liquid chromatography coupled with MS (LC–MS) analysis. The standard sample of phytic acid (P8810, Sigma-Aldrich) was dissolved in water to 10 mM first and then diluted to 75 nM for LC–MS.

The sample extracted from the purified TOC–TIC supercomplex preparation or the standard sample was analysed using high-performance LC (HPLC) (1290 Infinity, Agilent Technologies) coupled with a high-resolution mass spectrometer (6350 Accurate-Mass Q-TOF LC/MS, Agilent Technologies). The HPLC separation was performed on a C18 column (100 mm × 2.1 mm, Kinetex 2.6 μm C18 100 Å, Phenomenex) at a flow rate of 0.2 ml min^{-1} at 25 °C. The mobile phase consisted of two components, namely phase A (0.1% formic acid aqueous) and phase B (0.1% formic acid acetonitrile). The sample on the column was first eluted for 13 min with a linear gradient from 1% to 70% B and held for 4 min, followed by equilibration in 1% B. The MS analysis was performed in positive-ion mode (electrospray ionization (ESI)), with a scan range of 380–1,500 m/z for MS, or 20–1,500 m/z for MS/MS. Ion spray voltage was set at 3.5 kV in the positive mode, and the source temperature was 325 °C, with nebulizer gas set at 35 psi and flow rate at 5 l min^{-1}. In the product ion mode, the collision energy was set at 30 V. Analysis of the LC–MS data was performed using Peakview (v.2.1), and visualized using Matplotlib. The sample and standard had the same peak patterns of MS and MS/MS. Five main product ion peaks at an m/z of 773.85, 768.90, 751.56, 398.43 and 393.47 were detected, which may correspond to the metal complex of phytic acid and its degradation products.

Molecular dynamics simulation

The molecular dynamics simulations were performed on a truncated form of the TIC–TOC supercomplex with the bulk parts of TIC and ISC complexes within and nearby the inner membrane. The Toc75, Toc90, Toc34, Ctap4 subunits and chain X were manually removed from the model for the molecular dynamics simulation study. The peripheral regions of Tic100(82–237, 428–719 and 815–856), Ctap3(267–325 and 406–472), Tic214(532–700, 745–813, 1043–1492, 1667–1858 and 1971–1990) and Tic56(77–170) were also removed.

The model was embedded into a lipid bilayer sandwiched by ionized water boxes using the membrane builder function of CHARMM-GUI[58,59]. The lipid bilayer is composed of 40% monogalactosyl-diacylglycerol (MGDG, 16:0–18:1), 30% digalactosyl-diacylglycerol (DGDG, 16:0-18:1) and 30% 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) according to the previous lipid composition studies[60]. The lipids surrounded by the amphipathic helices of Ctap5 and Simp2 were modelled according to the electrostatic potential densities of their head group in Coot (v.0.9.8). The net charge of the system was neutralized by adding 150 mM KCl into the water box. The resulting system consists of 347,062 atoms, with an *xyz* dimension of 152 Å × 152 Å × 161 Å.

The simulations were performed using NAMD (v.2.14)[61] with the CHARMM36m force field. The particle mesh Ewald (PME) method was used to calculate long-range electrostatic interactions every 1 fs, with a grid density of 1 $Å^{-1}$ and a PME interpolation order of 6. A smoothing function was used for short-range non-bonded van der Waals and electrostatic interactions at a distance of 10 Å with a cut-off of 12 Å. The simulation was run using the NPT ensemble with the temperature set to 296 K and the pressure set to 1 atm. The simulations were first subjected to a minimization step, with a timestep of 1 fs for a total of 10,000 steps. For the subsequent equilibration step, the timestep was set to 1 fs for the first 385,000 steps and 2 fs for the next 750,000 steps. The final production step was run with a timestep of 1 fs.

To investigate the potential dynamic interactions of the transit peptide of the preprotein with the TIC–ISC complex, we docked the NMR structure of the *C. reinhardtii* preFdx1 transit peptide[62] (PDB: 1FCT) to the surface of the model from the aforementioned simulation result (output at 300 ns) using Autodock Vina (v.1.2.0)[63]. The bonds within the transit peptide were set as rotatable in AutodockTools (v.1.5.7) to search for the possible conformations, except for the planar peptide bonds and guanidinium bonds. The conformation with the lowest energy was appended to the molecular dynamics simulated model at 300 ns in VMD (v.1.9.4)[64], and the combined model was subjected to NAMD simulations with the same parameters as above.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The cryo-EM maps of the *C. reinhardtii* TOC–TIC supercomplex and the TOC complex, and their corresponding atomic coordinates, have been deposited in the Electron Microscopy Data Bank and the PDB under accession codes EMD-33528 and 7XZI, and EMD-33529 and 7XZJ, respectively. The other PDB files used for analysis and comparison in this work were downloaded from the RCSB PDB website (http://www.rcsb.org/), including the structures of *A. thaliana* Toc75 POTRA domains (5UAY), the NMR structure of the *C. reinhardtii* preFdx1 transit peptide (1FCT) and the crystal structure of RopP2 from *Ralstonia solanacearum* (5W3X). The full versions of all gel images are provided in Supplementary Data 1. All data analysed during this study are included in the Article and its Supplementary Information. Source data are provided with this paper.

45. Kropat, J. et al. A revised mineral nutrient supplement increases biomass and growth rate in *Chlamydomonas reinhardtii*. *Plant J.* **66**, 770–780 (2011).
46. Wittig, I., Braun, H. P. & Schägger, H. Blue native PAGE. *Nat. Protoc.* **1**, 418–428 (2006).
47. Wu, C., Huang, X., Cheng, J., Zhu, D. & Zhang, X. High-quality, high-throughput cryo-electron microscopy data collection via beam tilt and astigmatism-free beam-image shift. *J. Struct. Biol.* **208**, 107396 (2019).
48. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
49. Punjani, A., Zhang, H. & Fleet, D. J. Non-uniform refinement: adaptive regularization improves single-particle cryo-EM reconstruction. *Nat. Methods* **17**, 1214–1221 (2020).
50. Kimanius, D., Dong, L., Sharov, G., Nakane, T. & Scheres, S. H. W. New tools for automated cryo-EM single-particle analysis in RELION-4.0. *Biochem. J.* **478**, 4169–4185 (2021).
51. Pettersen, E. F. et al. UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).
52. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
53. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
54. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
55. Moriarty, N. W., Grosse-Kunstleve, R. W. & Adams, P. D. Electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation. *Acta Crystallogr. D* **65**, 1074–1080 (2009).
56. Afonine, P. V. et al. Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. D* **74**, 531–544 (2018).
57. Jurcik, A. et al. CAVER Analyst 2.0: analysis and visualization of channels and tunnels in protein structures and molecular dynamics trajectories. *Bioinformatics* **34**, 3586–3588 (2018).
58. Wu, E. L. et al. CHARMM-GUI Membrane Builder toward realistic biological membrane simulations. *J. Comput. Chem.* **35**, 1997–2004 (2014).
59. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem.* **29**, 1859–1865 (2008).
60. Mendiola-Morgenthaler, L., Eichenberger, W. & Boschetti, A. Isolation of chloroplast envelopes from *Chlamydomonas*. Lipid and polypeptide composition. *Plant Sci.* **41**, 97–104 (1985).
61. Phillips, J. C. et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* **153**, 044130 (2020).
62. Lancelin, J.-M. et al. NMR structures of ferredoxin chloroplastic transit peptide from *Chlamydomonas reinhardtii* promoted by trifluoroethanol in aqueous solution. *FEBS Lett.* **343**, 261–266 (1994).
63. Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. AutoDock Vina 1.2.0: new docking methods, expanded force field, and Python bindings. *J. Chem. Inf. Model.* **61**, 3891–3898 (2021).
64. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).

Acknowledgements We thank X.-J. Huang, B.-L. Zhu, X.-J. Li, D.-Y. Fan, L.-L. Zhang and the other staff members at the Center for Biological Imaging (CBI), Core Facilities for Protein Science at the Institute of Biophysics, Chinese Academy of Science (IBP, CAS) for the support in cryo-EM data collection; Z.-S. Xie, X. Ding, L.-L. Niu and J.-F. Wang at the CAS Research Platform for Protein Sciences for support with MS data collection and analysis; X.-B. Liang, Y.-X. Liu and X.-Y. Liu for technical support with biochemistry, cell culture and sample preparation; J.-Z. Lou and Y. Z. for their help in molecular dynamics simulations; S. Ramundo for sharing the HA-tagged Tic214 strain of *C. reinhardtii*; and W. Yang and N. Wang for recovering and aliquoting the cells. This work was funded by the Chinese Academy of Sciences Project for Young Scientists in Basic Research (YSBR-015), the National Natural Science Foundation of China (31925024), the Strategic Priority Research Program of CAS (XDB37020101 to Z.L.) and the National Key R&D Programme of China (2017YFA0503702).

Author contributions H.L. expressed and purified the protein complex sample, prepared the sample for the cryo-EM study, carried out cryo-EM data collection and prepared the figures. H.L. and A.L. processed the cryo-EM data. Z.L. and H.L. built and refined the atomic model, analysed the structure. H.L. designed the MS study and prepared the sample for the MS analysis. A.L. performed MD simulations on the TIC complex. Z.L., H.L. and A.L. wrote the initial draft and all of the authors edited the manuscript. J.-D.R. was involved in designing the project. Z.L. conceived and coordinated the research project.

Competing interests The authors declare no competing interests.

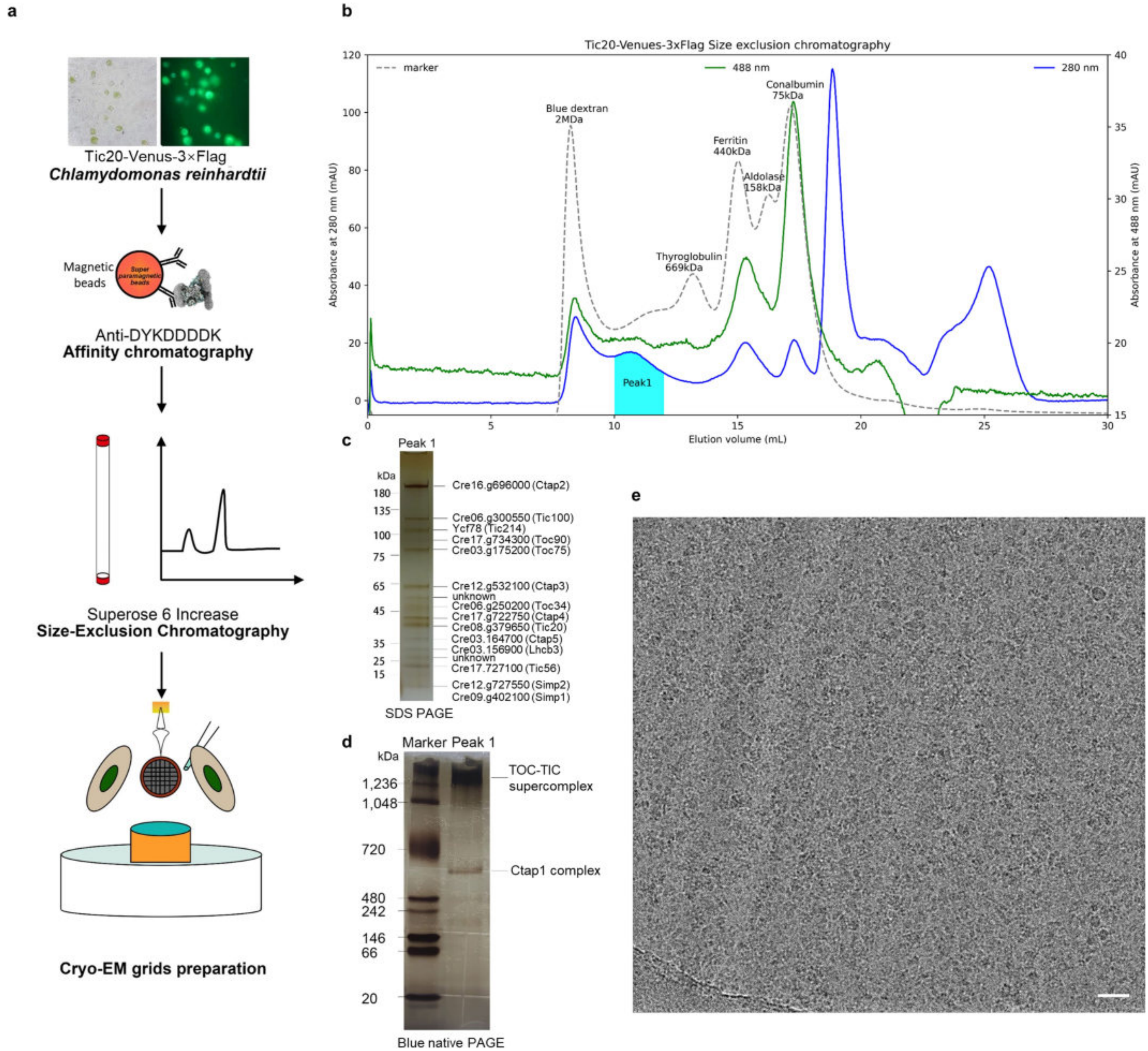
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-05744-y.

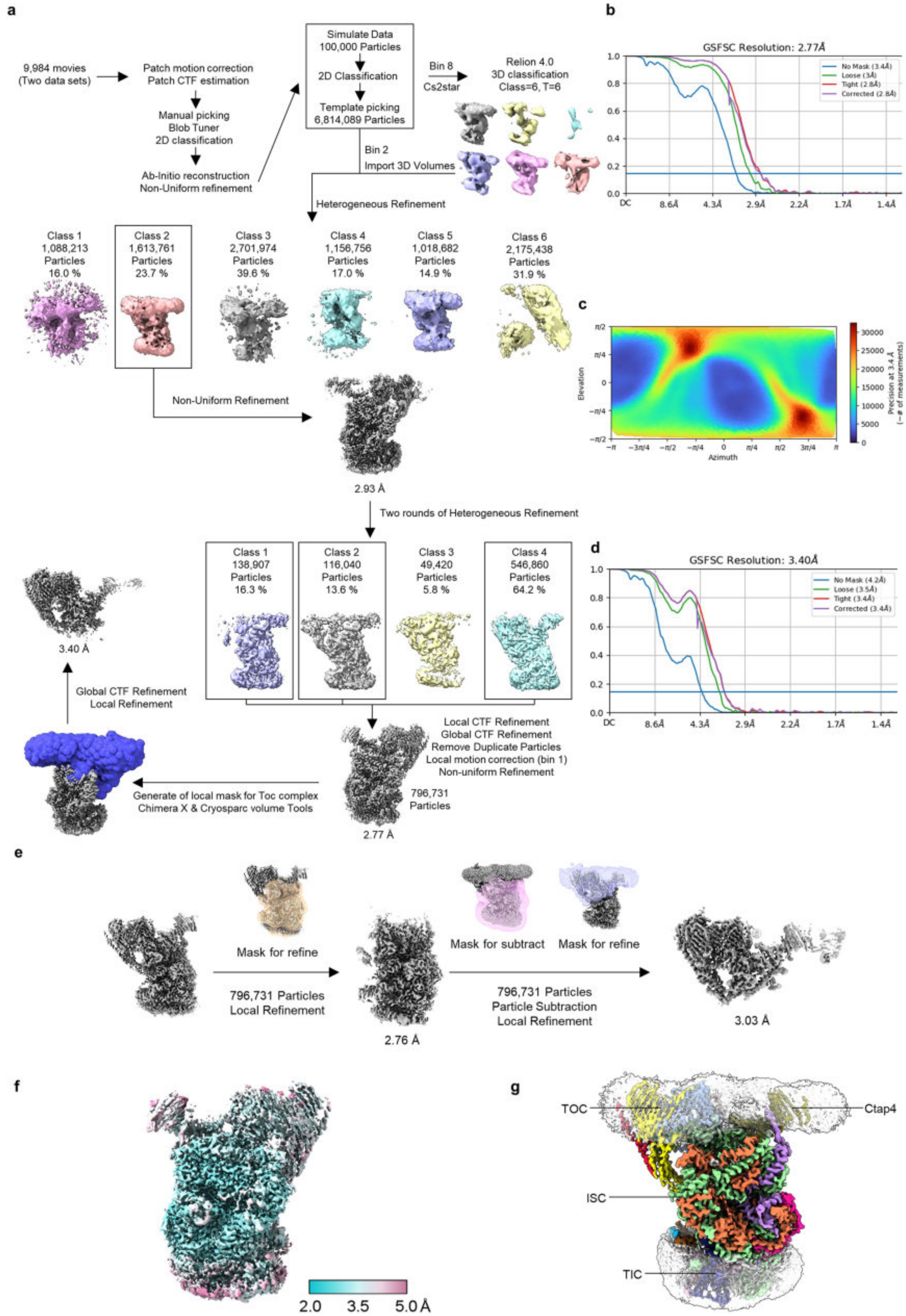
Correspondence and requests for materials should be addressed to Zhenfeng Liu.

Peer review information *Nature* thanks Alexey Amunts and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at http://www.nature.com/reprints.

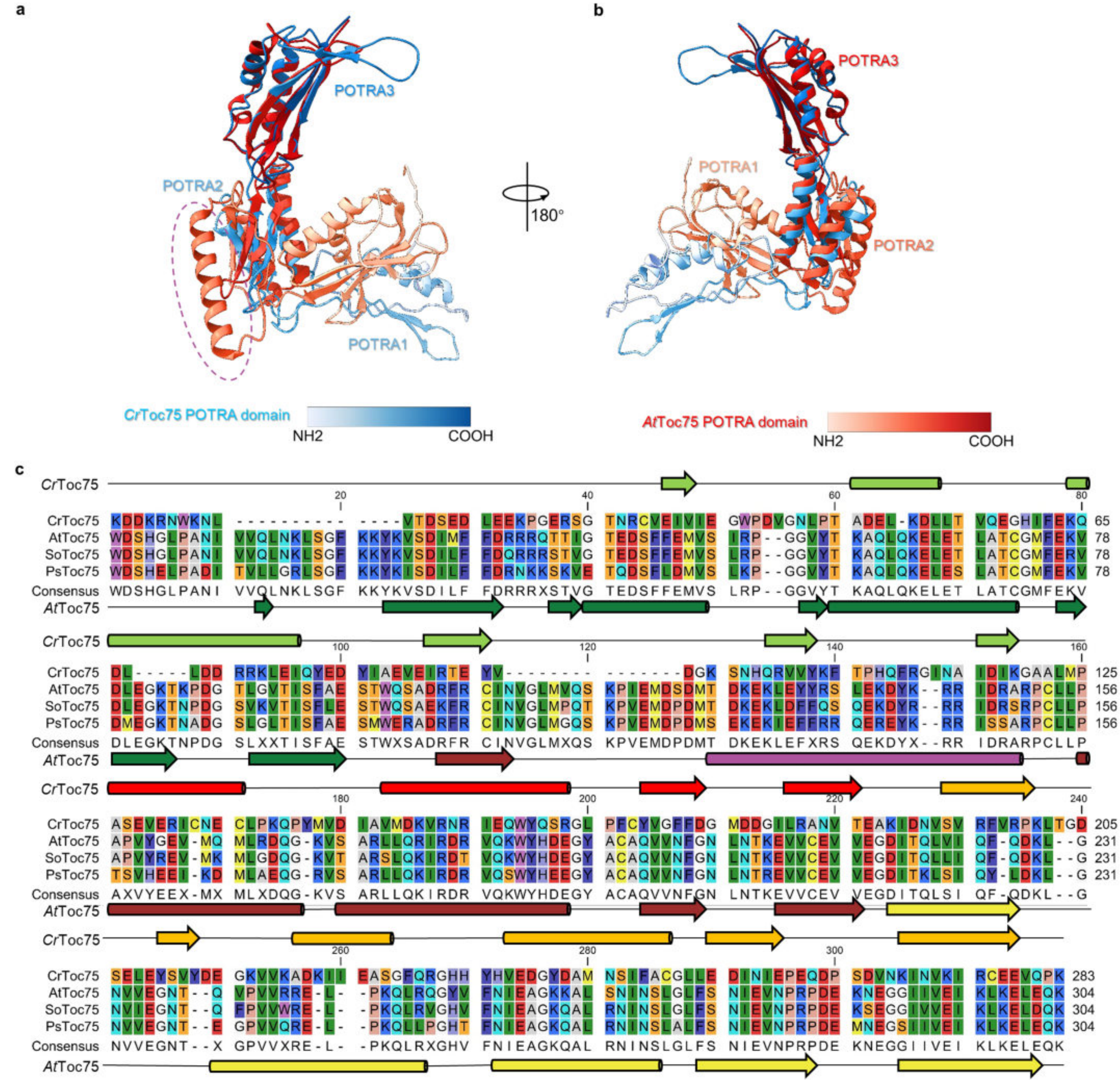


Extended Data Fig. 1 | Purification and characterization of the TOC-TIC supercomplex from *C. reinhardtii*. **a**, A schematic diagram illustrating the overall flowchart of the protein purification and cryo-EM grid preparation protocol. **b**, The profiles of size exclusion chromatography loaded with the sample eluted from the Anti-DYKDDDDK beads. The peak 1 fractions marked in the cyan area were pooled, concentrated and used for cryo-EM grid preparation. The blue solid line and green solid lines indicate the absorbance measured with the UV-Vis detector at 280 nm or 488 nm (the values and units are labelled on the Y axes on left and right sides) respectively. The grey dashed line is the size exclusion chromatography of the Gel Filtration Cal Kit High Molecular Weight sample (28-4038-42, Cytiva) loaded as a reference. **c** and **d**, The SDS-PAGE (c) and blue-native PAGE (d) analyses on the peak 1 fraction from the size exclusion chromatography. Silver staining of the gel was applied for visualization of the proteins. The identities of the protein bands are labelled on the right side according to the identification results obtained through mass spectrometry on the in-gel tryptic digestion products. Note that one of the FtsH-like AAA proteins named Ctap1 is present in the sample (d) and might be loosely associated with the TIC complex, but became separated from the TOC-TIC supercomplex during the purification process. For gel source data, see Supplemental Data 1. The SDS-PAGE and blue-native PAGE experiments were independently repeated three times and twice respectively with similar results. **e**, A representative cryo-EM micrograph of the TOC-TIC supercomplex. Scale bar, 20 nm. A total of 9,939 independent images of similar quality were collected from one cryo-EM grid and processed further.

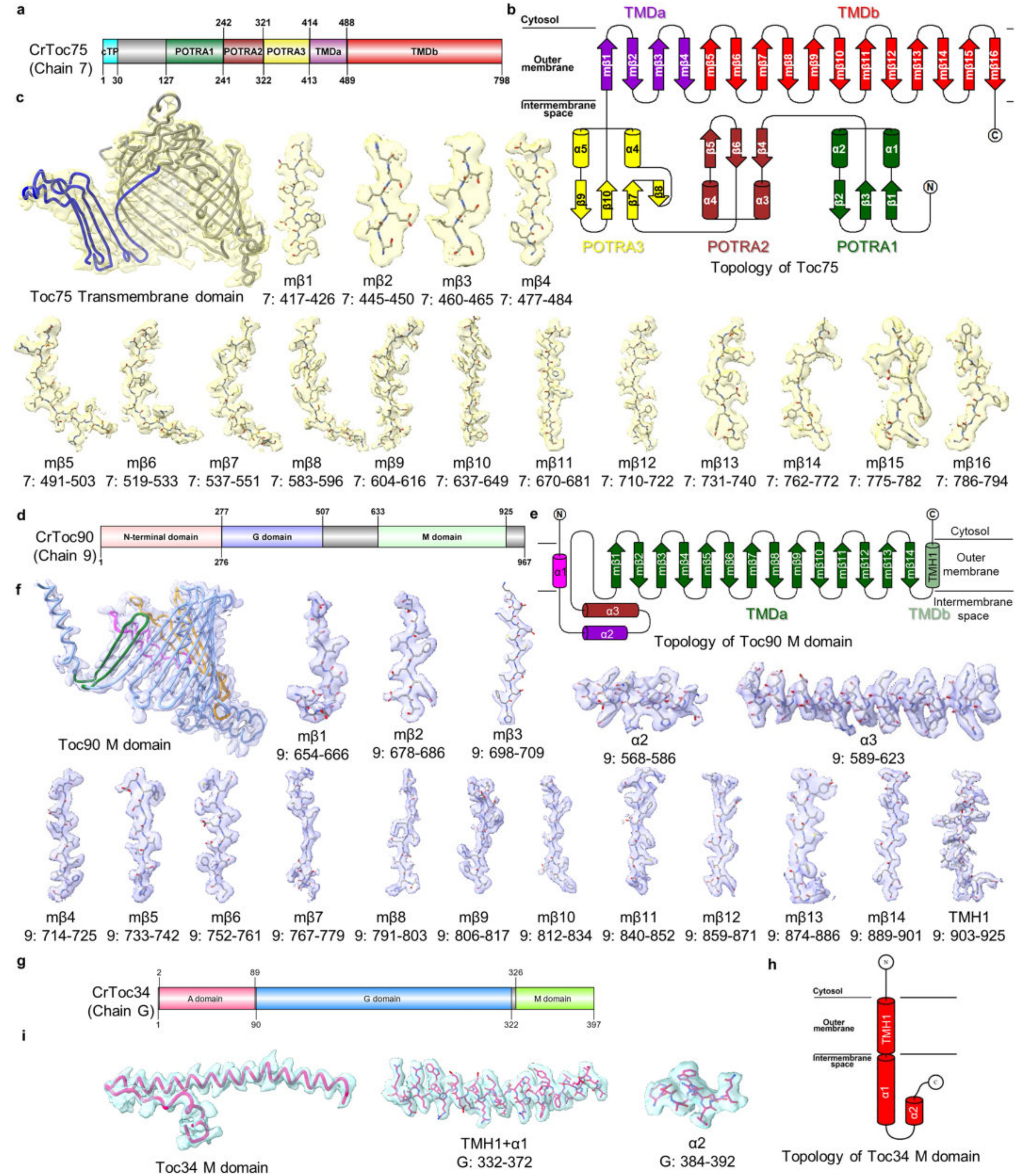


Extended Data Fig. 2 | See next page for caption.

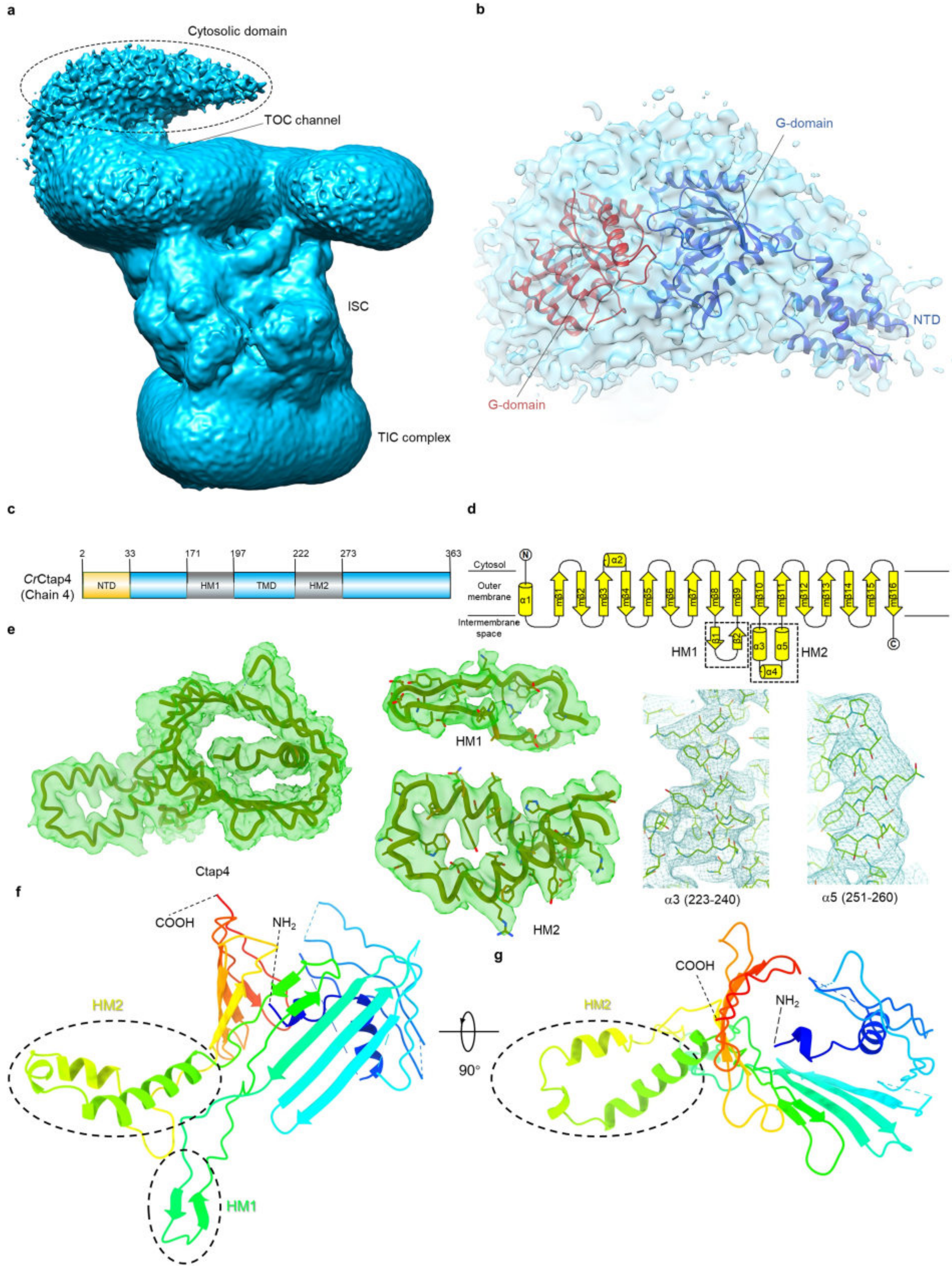
Extended Data Fig. 2 | Cryo-EM data processing of the TOC-TIC supercomplex. a, Flowchart of the cryo-EM data processing and analysis of the TOC-TIC supercomplex. (Please refer to the 'Cryo-EM data processing and refinement' section in Methods for more technical details). The two datasets were collected at different spots of the same grid sample. **b** and **d** The gold-standard Fourier shell correlation (GSFSC) curves of the TOC-TIC supercomplex map (b) and the local map around the TOC complex. **c**, The posterior precision directional distribution plot of the particles of the final reconstruction generated by cryoSPARC. **e**, The scheme for improving the densities around the TOC complex by applying the particle subtraction and local refinement procedure. **f**, The local resolutions for different regions of the final overall reconstruction of the TOC-TIC supercomplex. The local resolutions were estimated by cryoSPARC and visualized in ChimeraX. **g**, The densities of the detergent micelles surrounding the transmembrane domains of the TOC complex, Ctap4 and the TIC complex. The protein subunits are coloured in the same way as in Fig. 1, while the detergent micelles are in transparent white.



Extended Data Fig. 3 | Comparison of the POTRA domains of *C. reinhardtii* Toc75 with the corresponding ones from plant Toc75. a and **b**, Superposition of the cryo-EM structure of *Cr*Toc75 POTRA domains with the crystal structure of *Arabidopsis* Toc75 POTRA domains (PDB: 5UAY). The superposition was generated by using Chimera X matchmaker. **c**, Alignment of the amino acid sequences of the POTRA domains of Toc75 from *Chlamydomonas reinhardtii*, *Arabidopsis thaliana*, *Spinacea oleracea* and *Pisum sativum*. The sequence alignment was carried out by using the CLC Sequence Viewer 8. The β strands and α helices are represented by the arrows and cylinders respectively. Note that *At*Toc75 has an α-helix (marked in purple dotted circle in a and the purple cylinder in c) after the first β sheet of POTRA 2, which is not found in *Cr*Toc75 and has a sequence largely different from the corresponding region in *Cr*Toc75.

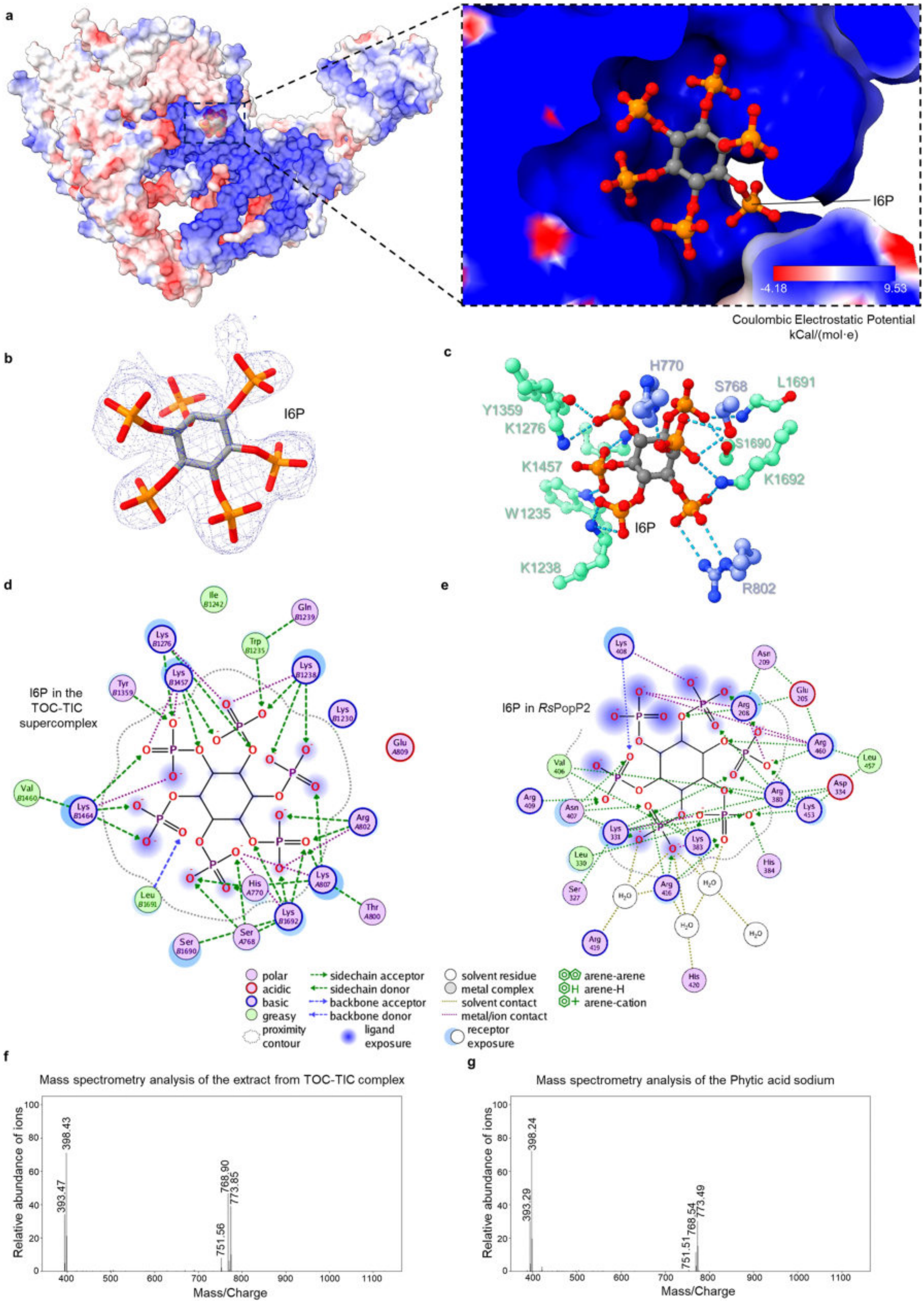


Extended Data Fig. 4 | Domain organization, secondary structure topology and cryo-EM densities of Toc75, Toc90 and Toc34. a, Domain organization of Toc75 protein. **b**, Topology of the secondary structures in Toc75. **c**, Cryo-EM densities of various parts in Toc75. The protein is built as chain 7 in the structural model and labelled as 7. The pore-lining mβ1-mβ5 region is highlighted in blue. **d**, Domain organization of Toc90 protein. **e**, Topology of the secondary structures in Toc90. **f**, Cryo-EM densities of various parts in Toc90. Chain 9 in the structural model corresponds to Toc90 (labelled as 9). The pore-lining α3-mβ1 loop, mβ1-mβ4, mβ13-mβ14 regions are highlighted in magenta, orange and green. **g**, Domain organization of Toc34 protein. **h**, Topology of the secondary structures in Toc34. **i**, Cryo-EM densities of various parts in Toc34. Chain G in the structural model corresponds to Toc34 (labelled as G).



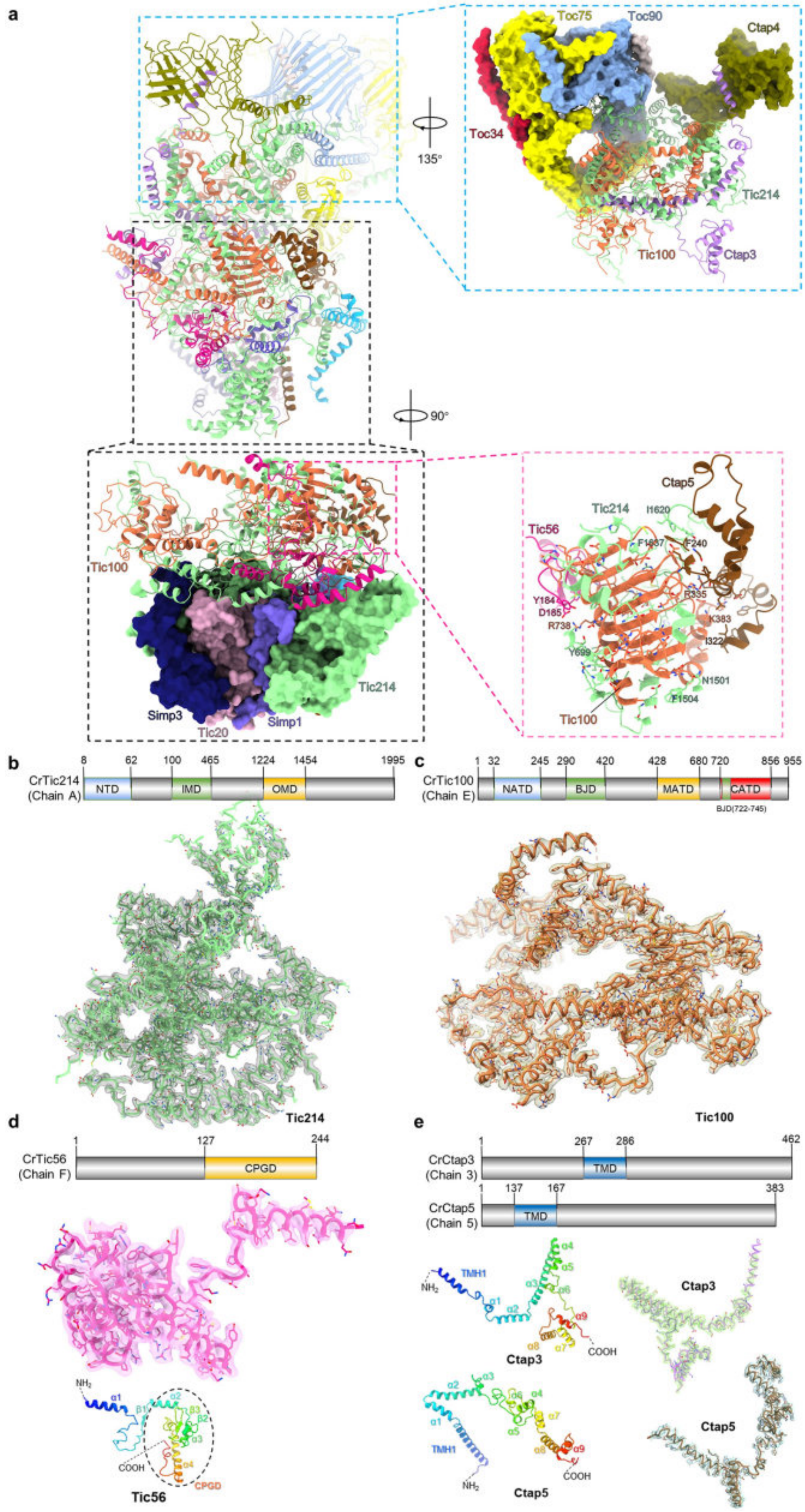
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | The cryo-EM densities in the cytoplasmic region above the TOC complex and the structure of Ctap4. a, The cryo-EM map of TOC-TIC supercomplex filtered through the Gaussian filter. The contour level is at 0.011 (Chimera). Apparently, there are some weak cryo-EM densities above the TMDs of the Toc75-Toc90-Toc34 complex, potentially belonging to the G domains of Toc90 and Toc34. The densities inside the elliptical dashed ring may contain the putative G-domains of Toc90 and Toc34. The G domains are likely very mobile as the densities are much weaker than the transmembrane domains and the loops connecting them to the transmembrane domains are barely visible in the map. **b**, Fitting of the cytosolic domain densities with a heterodimeric model of the G domains from Toc34 and Toc90. The model was constructed by referring to the homodimeric structure of Toc34 G domain (PDB code: 3BB1). The map contour level is at 0.0125. Colour codes: red, G-domain of Toc34; blue, G-domain and the N-terminal domain (NTD) of Toc90. **c**, Domain organization of the Ctap4 protein. **d**, Topological arrangement of the secondary structures in Ctap4 as predicted by Alphafold2. **e**, Cryo-EM densities of Ctap4 (chain 4). **f** and **g**, Side view and top view of Ctap4 structure. HM1 and HM2, hydrophilic motifs 1&2.



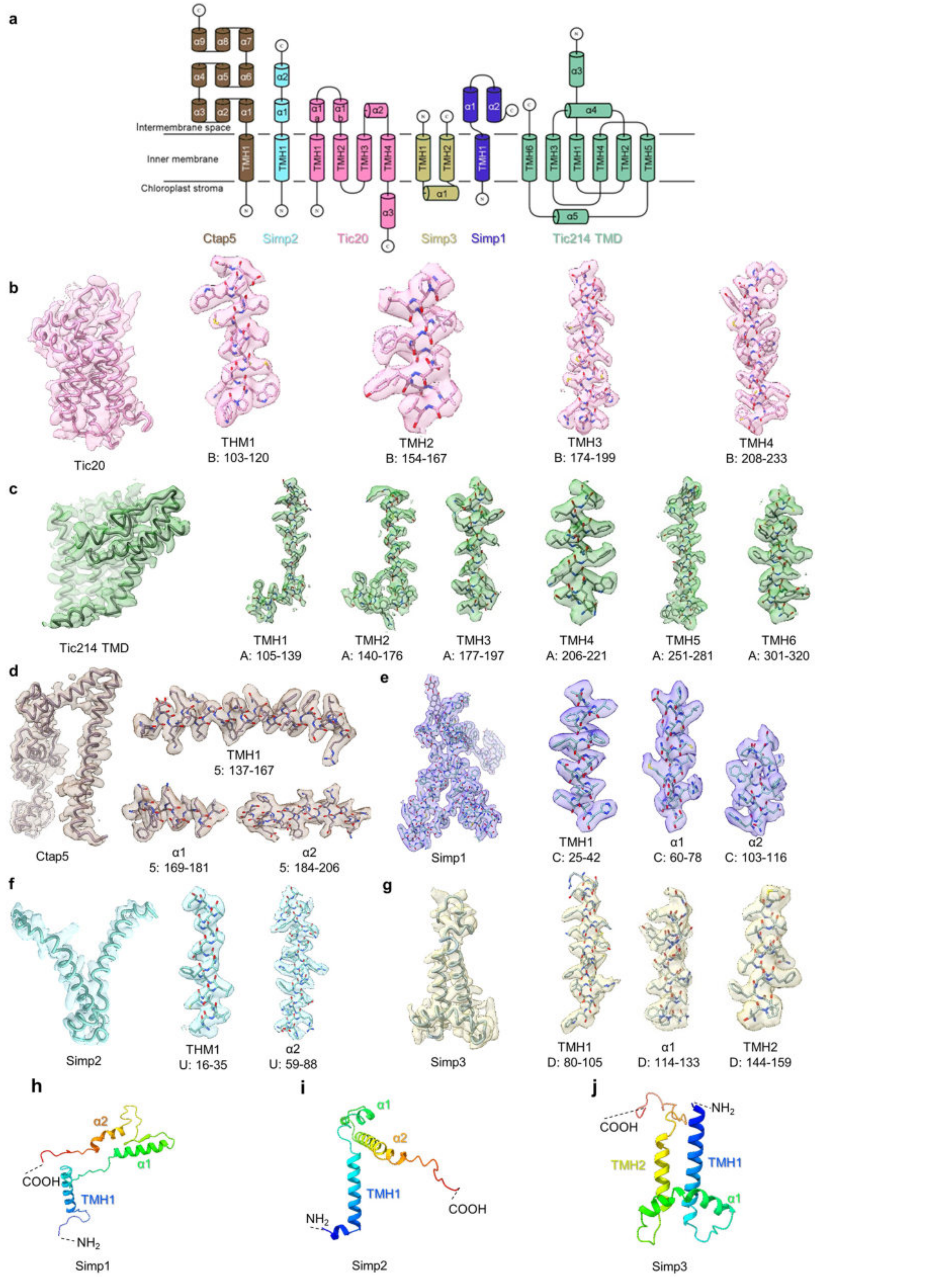
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Location and identification of an $InsP_6$ molecule in the TOC-TIC supercomplex from *C. reinhardtii*. **a**, Electrostatic surface representation of the TOC complex hosting an $InsP_6$-binding site. The zoom-in view on the right side shows an $InsP_6$ molecule accommodated in a positive charge-rich cavity. The electropositive surface of the cavity is complementary to the negative charges carried by the phosphate groups of $InsP_6$. **b**, The cryo-EM density of the $InsP_6$ molecule with the stick model superposed. The cryo-EM map is contoured at 0.25 V. **c**, Interactions of $InsP_6$ with adjacent amino acid residues from Tic214 (green) and Toc90 (blue) in the TOC-TIC supercomplex. **d and e**, The 2D interaction map of $InsP_6$ in the TOC-TIC supercomplex (d) or *Rs*PopP2 (PDB:5W3X, e) calculated by using the MOE program. *Rs*PopP2 is a bacterial effector protein produced by the plant pathogen *Ralstonia solanacearum* and hosts an $InsP_6$-binding site similar to the one found in the TOC complex. The phosphate groups of $InsP_6$ molecules are surrounded by positively-charged residues in both cases. **f** and **g**, Mass spectrometry analysis of the extract from the TOC-TIC supercomplex sample (f) or the standard sample of phytic acid (sodium salt, g).

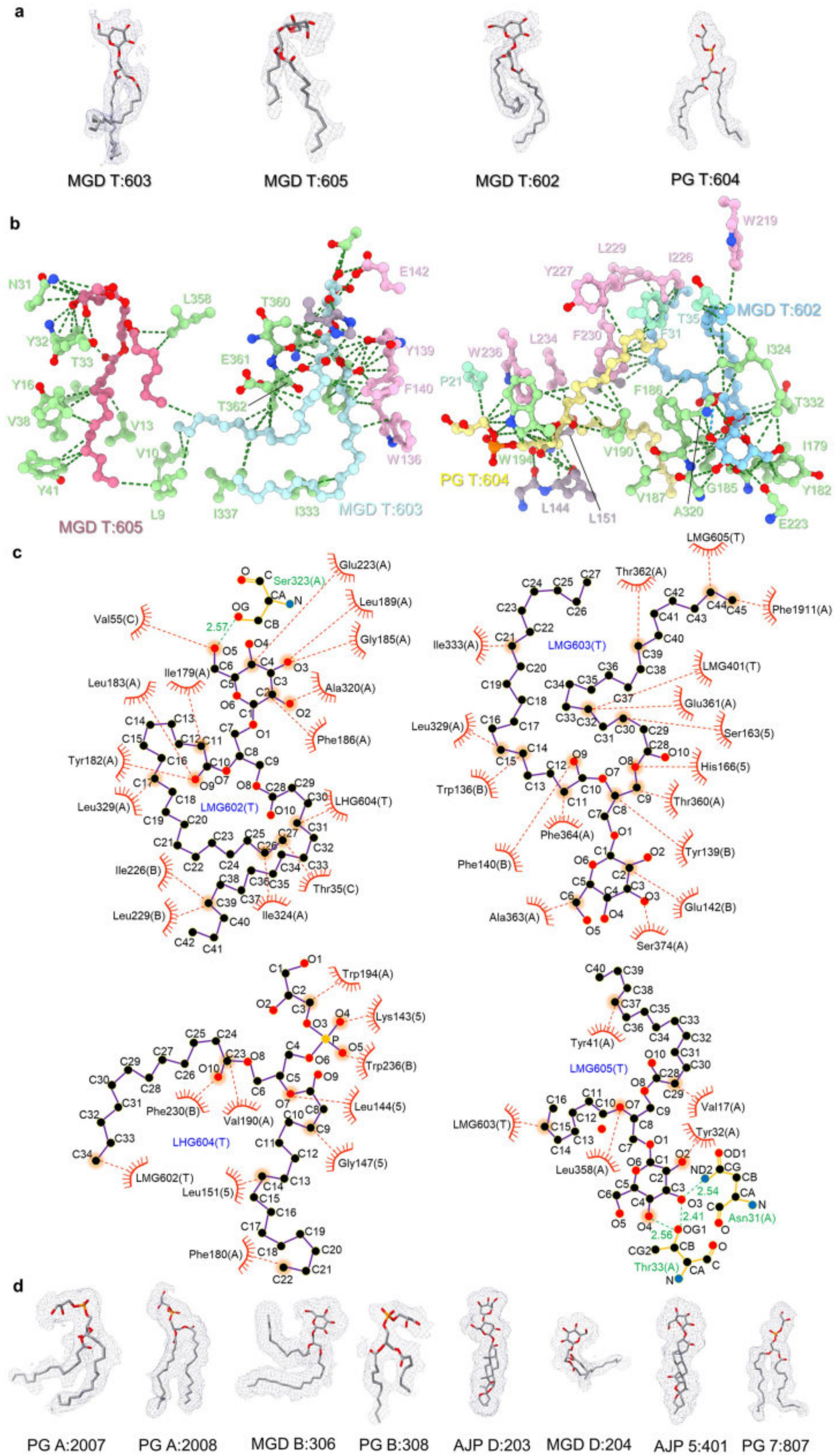


Extended Data Fig. 7 | See next page for caption.

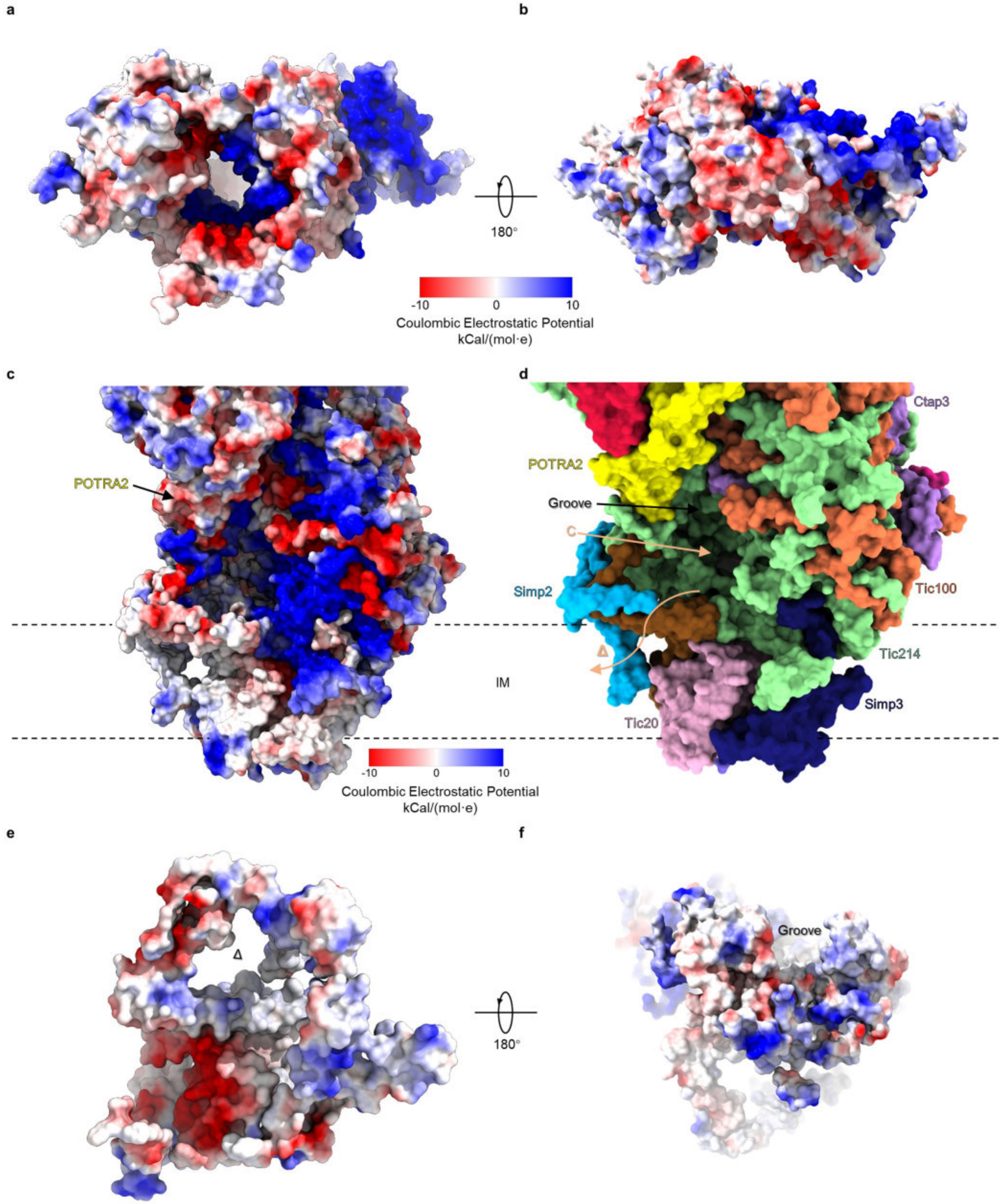
Extended Data Fig. 7 | The assembly and components of the intermembrane space complex in the TOC-TIC supercomplex. a, The assembly interfaces between ISC and the TOC/TIC complex. **b**, Domain organization and cryo-EM density of the intermembrane space domain of Tic214. **c**, Domain organization and cryo-EM density of Tic100. **d**, Domain organization, density and cartoon model of Tic56. **e**, Domain organization, densities and cartoon models of Ctap3 and Ctap5. Ctap3 has an extended α-helical domain (Arg295-Ser472) at the carboxy-proximal region, intertwining with the bowl-like region of Tic214-Tic100 complex. It docks two short α-helices (α1 and α2) on the outer surface of the upper half of ISC, inserts a long α-helix (α3) and two short ones (α4 and α5) into the bottom of the bowl-like structure formed by Tic214 and Tic100. The last three α-helices (α7-9) of Ctap3 bind to the amino-terminal helix of Tic56 and the carboxy-terminal helix of Tic100. On the opposite side, Ctap5 attaches its carboxy-proximal α-helical domain (Glu215-Ala329) to a surface groove of the Tic100-Tic214 complex below the POTRA1 and POTRA2 domains of Toc75. It interacts closely with the BJD of Tic100 and Tic214 (Gln1621-Phe1645 and Tyr1917-Ile1949) and provides binding sites for the POTRA1 and POTRA2 domains of Toc75. Thereby, Ctap5 may help to stabilize the Tic214-Tic100 complex and serves as a bridge to connect Toc75 with the inner-membrane complex.



Extended Data Fig. 8 | Topology and cryo-EM densities of the six proteins involved in forming the TIC complex. a, Topological arrangement of secondary structures in Ctap5 (chain 5), Simp2 (chain U), Tic214 (chain A), Tic20 (chain B), Simp3 (chain D) and Simp1 (chain C) proteins. **b**-**g**, Cryo-EM densities of various parts in the six proteins. **h**-**j**, Cartoon models of Simp1, Simp2 and Simp3.



Extended Data Fig. 9 | The lipid and detergent molecules associated with the TOC-TIC supercomplex. a, The cryo-EM densities of four central lipid molecules located in or above the putative pore of the TIC complex. **b**, Interactions of the four lipid molecules with adjacent amino acid residues. Green, Tic214; Pink, Tic20; brown, Ctap5. **c**, LigPlot analyses showing the hydrogen bonds and hydrophobic interactions formed between the four lipid molecules and nearby amino acid residues. **d**, The cryo-EM densities of other peripheral lipid molecules associated with Tic214(A), Tic20(B), Simp3(D), Ctap5(5) and Toc75(7).



Extended Data Fig. 10 | The surface presentations of the TOC complex, the intermembrane space domain connecting the TOC channel with the TIC channel, and the TIC complex. a and b, The electrostatic potential surface of the TOC complex viewed from the cytoplasmic and intermembrane space sides, respectively. **c**, The electrostatic potential surface of the ISC connecting the TOC complex with the TIC complex. **d**, The presence of a surface groove connecting the channels of the TOC and TIC complexes. The local regions labelled by "C" and "Δ" are the cavity above the central funnel of the TIC complex and the triangular entrance formed by Ctap5, Simp2 and Tic214, respectively. **e and f**, The electrostatic potential surface of the TIC complex viewed from the intermembrane space and stromal sides, respectively.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	#1 TOC-TIC supercomplex (EMDB-33528, PDB 7XZI)	#2 TOC complex (local) (EMDB-33529, PDB 7XZJ)
Data collection and processing		
EM equipment	FEI Titan Krios G3	
Detector	Gatan K2	
Magnification	105,000×	105,000×
Voltage (kV)	300	300
Electron exposure (e–/Å^2)	50	50
Number of frames/movie	33	33
Defocus range (μm)	-1.5~1.8	-1.5~1.8
Pixel size (Å)	0.675	0.675
Reconstruction		
Software	Cryosparc v3.3 & Relion 4.0	
Symmetry imposed	C1	C1
Initial particle images (no.)	6,814,089	6,814,089
Final particle images (no.)	796,731	796,731
Map resolution (Å)	2.77	3.03
FSC threshold	0.143	0.143
Model building		
Software	WinCoot 0.9.4 & AlphaFold2	
Refinement		
Software	PHENIX 1.20.1-4487	
Model resolution (Å)	2.8/3.2	3.4/4.0
FSC threshold	0.143/0.5	0.143/0.5
Map sharpening *B* factor (Å^2)	-85.6	-63.8
Model composition		
Non-hydrogen atoms	37,362	13,900
Protein residues	4,771	1,890
Lipids/Detergents/other ligand	10/5/1	0/0/1
B factors (Å^2)		
Protein	108.71	109.12
Ligand	121.25	142.20
R.m.s. deviations		
Bond lengths (Å)	0.002	0.011
Bond angles (°)	0.493	1.339
Validation		
MolProbity score	1.29	1.45
Clashscore	3.57	3.62
Poor rotamers (%)	0.85	0.40
Ramachandran plot		
Favored (%)	97.20	95.77
Allowed (%)	2.80	4.23
Disallowed (%)	0.00	0.00

See also Extended Data Fig. 2 for more details on cryo-EM data processing, refinement and evaluation.

nature portfolio

Corresponding author(s): Zhenfeng Liu

Last updated by author(s): Dec 29, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	SerialEM v3.6/3.6.1
Data analysis	RELION 4.0, cryoSPARC v3.2, UCSF Chimera 1.16 and ChimeraX 1.3/1.4, WinCoot 0.9.4/0.9.8, PHENIX 1.20.1-4487, CAVER Analyst 2.0, NAMD-2.14, Autodock Vina v.1.2.0, VMD 1.9.4, Peakview 2.1, MASCOT 2022, Thermo Proteome Discoverer program (1.4.0.288), Matplotlib 1.16.0, Python 3.7.10, NAMD 2.14, AutodockTools 1.5.7, VMD 1.9.4, CHARMM-GUI 2022

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The cryo-EM maps of the Chlamydomonas reinhardtii TOC-TIC supercomplex and the TOC complex, and their corresponding atomic coordinates have been deposited in the Electron Microscopy Data Bank and the Protein Data Bank under the accession codes EMD-33528 and 7XZI, EMD-33529 and 7XZJ respectively. The

other PDB files used for analysis and comparison in this work were downloaded from the RCSB PDB website (http://www.rcsb.org/), including the structures of Arabidopsis thaliana Toc75 POTRA domains [accession code: 5UAY], the NMR structure of the C. reinhardtii preFdx1 transit peptide (accession code: 1FCT) and the crystal structure of RopP2 from Ralstonia solanacearum (accession code: 5W3X). The full versions of all gel images are provided in Supplementary Figure 1. The source data for Fig. 5b&g and Extended Data Fig. 1b are provided in Supplementary Tables 4 and 5 respectively. All data analyzed during this study are included in this Article and its Supplementary Information.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Statistical methods were not used to predetermine the sample size. A total of 9,984 micrographs (two data sets) were collected from one cryo-EM grid with thin sample loaded on thousands of tiny holes. Usually, particles extracted from a dataset of 2000-3000 micrographs of good quality should be sufficient for obtaining a reconstruction at near atomic resolution. As the number of TOC-TIC supercomplex particles per micrograph is relatively small (usually less than 100 particles), more micrographs were collected to obtain sufficient particle images for reconstruction of a high-resolution map. After particle picking, 6,814,089 raw single particle images were extracted from the micrographs and used for further processing. No statistical methods were used to determine sample size reconstruction of the TOC-TIC supercomplex. A total of 796,731 particle images of good quality were selected through 2D classification and 3D classification procedures. The sample size of TOC-TIC supercomplex particles is much larger than the standard sample sizes (100-300k) for reconstruction of other membrane proteins.
Data exclusions	The initial cryo-EM micrographs are screened manually to exclude those with weak signals or broken sample, which is a standard procedure for cryo-EM data processing. After screening, a total of 45 micrographs were excluded due to low quality and 9939 micrographs of good quality were kept for further processing.
Replication	Two datasets collected on one cryo-EM grid were combined for further processing. Structure determination was carried out once. Biochemical experiments were repeated independently from 2 to 4 times, and were all successfully reproduced. The numbers of independent experiments and replicates are indicated in the related figure legends.
Randomization	Randomization is not relevant for the structural analysis in our study because no grouping is needed and the results are not subjective. For single-particle cryo-EM analysis, the particle images are divided into two random sets for reconstruction, refinement and estimation of the resolution via the Gold-Standard Fourier Shell Correlation (FSC) method.
Blinding	Blinding is not relevant for the structural analysis in our study because no allocation into experimental groups was needed and the samples under investigation are required to be known so that the investigators can select the correct target particles for reconstruction and further analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Matters arising

Multivariate BWAS can be replicable with moderate sample sizes

https://doi.org/10.1038/s41586-023-05745-x

Received: 10 April 2022

Accepted: 19 January 2023

Published online: 8 March 2023

Open access

Check for updates

Tamas Spisak[1,2]✉, Ulrike Bingel[2] & Tor D. Wager[3]

ARISING FROM: S. Marek et al. Reproducible brain-wide association studies require thousands of individuals. *Nature* https://doi.org/10.1038/s41586-022-04492-9 (2022)

Brain-wide association studies (BWAS)—which correlate individual differences in phenotypic traits with measures of brain structure and function—have become a dominant method for linking mind and brain over the past 30 years. Univariate BWAS typically test tens to hundreds of thousands of brain voxels individually, whereas multivariate BWAS integrate signals across brain regions into a predictive model. Numerous problems have been raised with univariate BWAS, including a lack of power and reliability and an inability to account for pattern-level information embedded in distributed neural circuits[1–4]. Multivariate predictive models address many of these concerns, and offer substantial promise for delivering brain-based measures of behavioural and clinical states and traits[2,3].

In their recent paper[4], Marek et al. evaluated the effects of sample size on univariate and multivariate BWAS in three large-scale neuroimaging datasets and came to the general conclusion that "BWAS reproducibility requires samples with thousands of individuals". We applaud their comprehensive analysis, and we agree that (1) large samples are needed when conducting univariate BWAS and (2) multivariate BWAS reveal substantially larger effects and are therefore more highly powered.

Marek et al.[4] find that multivariate BWAS provide inflated in-sample associations that often cannot be replicated (that is, are underpowered) unless thousands of participants are included. This implies that effect-size estimates from the discovery sample are necessarily inflated. However, we distinguish between the effect-size estimation method (in-sample versus cross-validated) and the sample (discovery versus replication), and show that, with appropriate cross-validation, the in-sample inflation that Marek et al.[4] report in the discovery sample can be entirely eliminated. With additional analyses, we demonstrate that multivariate BWAS effects in high-quality datasets can be replicable with substantially smaller sample sizes in some cases. Specifically, applying a standard multivariate prediction algorithm to functional connectivity in the Human Connectome Project yielded replicable effects with sample sizes of 75–500 for 5 of 6 phenotypes tested (Fig. 1).

These analyses are limited to a selected number of phenotypes in a relatively high-quality dataset (measured in a young adult population with a single scanner) and should not be overgeneralized. However, they highlight that the key determinant of sample size requirements is the true effect size of the brain–phenotype relationship and that, with proper internal validation, appropriate effect-size estimates and sufficiently large effects for moderately sized studies are possible.

Marek et al.[4] evaluate in-sample effect-size inflation in multivariate BWAS by training various multivariate models in a 'discovery sample' and comparing the in-sample effect sizes (prediction–outcome correlation, r) estimated from the training sample to the performance in an independent replication sample. On the basis of a bootstrap analysis, with variously sized pairs of samples drawn randomly from the Adolescent Brain Cognitive Development study, the authors report a severe effect-size inflation of $\Delta r = -0.29$ (average difference between the in-sample effect sizes in the discovery sample and the out-of-sample effect sizes in the replication sample) and conclude that "[e]ven at the largest sample sizes ($n \approx 2{,}000$), multivariate in-sample associations remained inflated on average".

The issue with claims of inflation is that the in-sample effect size estimates of Marek et al.[4] were based on training multivariate models on the entire discovery sample, without cross-validation or other internal validation (as confirmed by inspection of the code and discussion with the authors). Such in-sample correlations are not valid effect-size estimates, as they produce a well-known overfitting bias that increases with model complexity[5]. Standard practice in machine learning is to evaluate model accuracy (and other performance metrics) on data independent of those used for training. In line with current recommendations for multivariate brain–behaviour analyses[6,7], this is typically performed using internal cross-validation (for example, k-fold) to estimate unbiased effect sizes in a discovery sample, and (less commonly) further validating significant cross-validated effects in held-out or subsequently acquired replication samples[2,5].

Using cross-validation to estimate discovery-sample effects impacts the pool of studies selected for replication attempts, the degree of effect-size attenuation in replication samples, and the sample size needed for effective replication and mitigation of publication bias. To demonstrate this and provide quantitative estimates of sample size requirements in multivariate BWAS, we analysed functional connectivity data from the Human Connectome Project[8] (one of the datasets in Marek et al.[4]) using cross-validation to estimate discovery-sample effect sizes. As shown in Fig. 1a–d, cross-validated discovery effect-size estimates are unbiased (that is, not inflated on average), irrespective of the sample size and the magnitude of the effect. As expected, even with cross-validation, smaller sample sizes resulted in lower power (Fig. 1e) and increased variability in effect-size estimates across samples (Fig. 1c). Such variability is undesirable because it reduces the probability of independent replication (Fig. 1f). Moreover, selection biases—most notably, publication bias—can capitalize on such variability to inflate effect sizes in the literature (Fig. 1g). Although these effects of using small sample sizes are undesirable, they do not invalidate

[1]Institute of Diagnostic and Interventional Radiology and Neuroradiology, University Medicine Essen, Essen, Germany. [2]Center for Translational Neuro- and Behavioral Sciences, Department of Neurology, University Medicine Essen, Essen, Germany. [3]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA. ✉e-mail: tamas.spisak@uk-essen.de

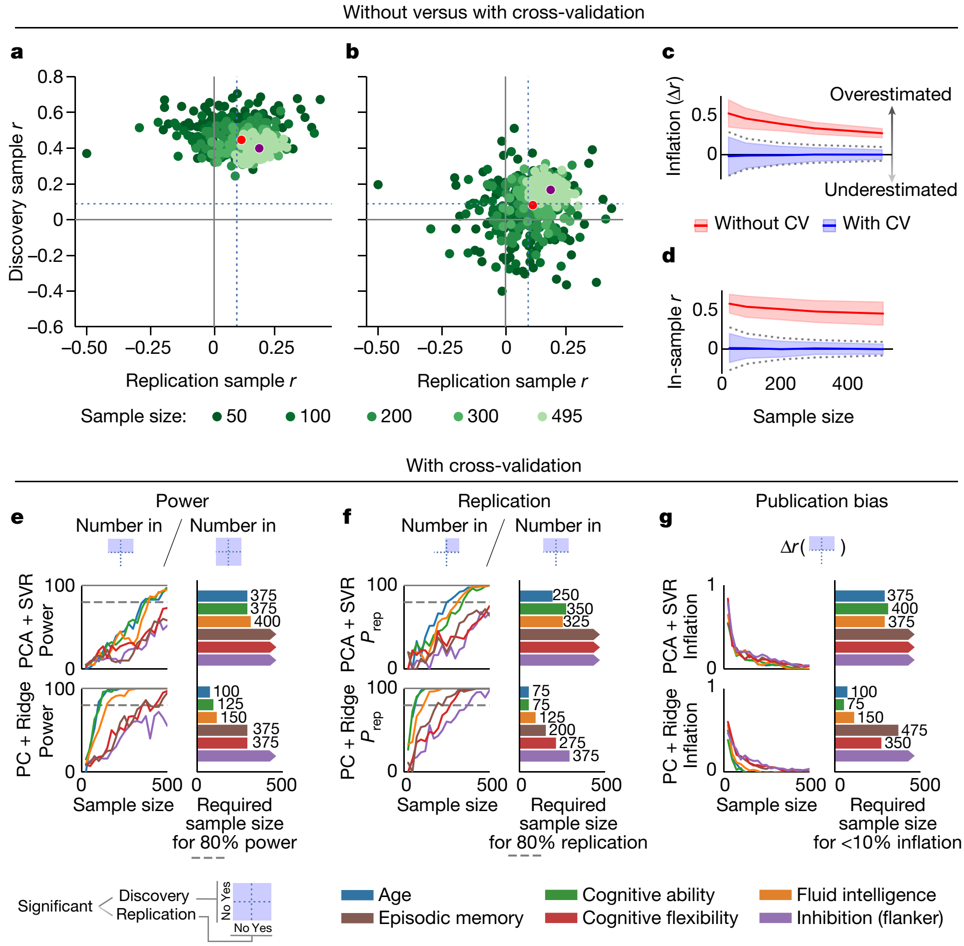


Fig. 1 | Examples of multivariate BWAS providing unbiased effect sizes and high replicability with low to moderate sample sizes. a, Discovery sample effects in multivariate BWAS are inflated only if estimates are obtained without cross-validation (CV). **b**, Cross-validation fully eliminates in-sample effect-size inflation and, as a consequence, provides higher replicability. Data are from the Human Connectome Project (HCP1200, PTN release, $n = 1{,}003$). Each point in **a** and **b** corresponds to one bootstrap subsample, as in figure 4b of Marek et al.[4]. The dotted lines denote the threshold for $P = 0.05$ with $n = 495$. Mean multivariate brain–behavioural phenotype associations across 100 bootstrap samples at $n = 200$ and for the full sample are denoted by red and purple dots. **c**, The inflation of in-sample effect size obtained without cross-validation (red) is reduced, but does not disappear, at higher sample sizes. Conversely, cross-validated estimates (blue) are slightly pessimistic with low sample sizes and become quickly unbiased as sample size is increased. **d**, Without cross-validation, in-sample effect-size estimates are non-zero ($r \approx 0.5$, red), even when predicting permuted outcome data. Cross-validation eliminates systematic bias across all sample sizes (blue). The dashed lines in **c** and **d** denote 95% parametric confidence intervals, and the shaded areas denote bootstrap- and permutation-based confidence intervals. **e**,**f**, Cross-validated analysis reveals that sufficient in-sample power (**e**) and out-of-sample replication probability (P_{rep}) (**f**) can be achieved for a variety of phenotypes at low or moderate sample sizes. 80% power and P_{rep} are achievable in <500 participants for 3 out of 6 phenotypes (coloured bars) using the prediction algorithm of Marek et al.[4] (**e** and **f** (top), the sample size required for 80% power or P_{rep} is shown). The remaining three phenotypes require sample sizes of >500 (bars with arrows). Power and P_{rep} can be substantially improved with a ridge regression-based model recommended in some comparison studies[10,11] (**e** and **f** (bottom), with 80% power and P_{rep} with sample sizes as low as $n = 100$ and $n = 75$, respectively, when predicting cognitive ability, and sample sizes between 75 and 375 for other investigated variables (fluid intelligence, episodic memory and cognitive flexibility), except inhibition assessed with the flanker task, which replicated with $n = 375$ but did not reach 80% power with $n = 500$. **g**, We estimated interactions between sample size and publication bias by computing effect size inflation ($r_{discovery} - r_{replication}$) only for those bootstrap cases in which prediction performance was significant ($P > 0.05$) in the replication sample. Our analysis shows that the effect-size inflation due to publication bias is modest (<10%) with fewer than 500 participants for half of the phenotypes using the model from Marek et al.[4] and all phenotypes but the flanker using the ridge model. The blue squares show conditional relationships assessed to derive metrics in **e**, **f** and **g** with reference to **b**. The top and bottom squares indicate positive and negative results in the discovery sample, respectively. The left and right squares indicate negative and positive results in the replication sample. The blue squares indicate how these conditions were applied to derive the metrics.

the use of multivariate BWAS in small samples, and publication biases can be mitigated by practices that, like internal cross-validation, are quickly becoming standards in the field[2,5]. These include preregistration, registered reports, reporting confidence intervals and the use of hold-out samples tested only once on a single, optimized model to avoid overfitting.

Given these considerations, we wondered how many participants are generally required for multivariate BWAS. The answer to this question depends on the reliability of both phenotypic and brain measures, the size of the effects linking them, the algorithm and model-selection steps used and the use cases for the resulting brain measures. For example, multivariate models trained on as few as 20 participants[9] can have high reliability (ICC = 0.84)[10], broad external validity and large effect sizes (Hedges $g = 2.3$)[11] in independent samples (for example, more than 600 participants from 20 independent studies in ref. [11]) when predicting behavioural states within-person rather than traits. In this case, the benefit of large samples is primarily in accurately estimating local brain weights (model parameters)[12] rather than increasing out-of-sample

accuracy. Here we performed functional connectivity-based multivariate BWAS with cognitive ability (the phenotype shown in figure 4 of Marek et al.[4]) and five other cognition-related example phenotypes selected at random and demonstrate that, even when predicting trait-level phenotypes, as Marek et al.[4] did, sample sizes of 75–500 are sufficient in five out of six of cases that we tested (or three out of six cases using the prediction algorithm of Marek et al.[4]) to achieve high statistical power and replicability (for example, 80%) and to mitigate effect size inflation due to publication bias.

The basis for these estimates is shown in Fig. 1e–g. Using cross-validated discovery sample effect-size estimates, the multivariate BWAS model of Marek et al.[4]—principal-component-based reduction of bivariate connectivity followed by support vector regression (PCA + SVR)—showed 80% in-sample power and 80% out-of-sample replication probability (P_{rep}) at $n < 500$ for three out of six phenotypes that we examined (age, cognitive ability and fluid intelligence). However, this model has been shown to be disadvantageous in some comparison studies[12,13]. We therefore performed the same power and sample-size calculations for a multivariate BWAS using another approach—ridge regression on partial correlation matrices with a default shrinkage parameter of 1 (PC + ridge; Supplementary Methods). Although this approach is still probably suboptimal[12,13] (we avoided testing other models to avoid overfitting), it substantially improved the power (Fig. 1e (bottom)), independent replication probability ([P_{rep}]; Fig. 1f (bottom)) and resistance to inflation due to publication bias (Fig. 1g (bottom)). Eighty per cent power and P_{rep} were achieved at sample sizes from 75 to 150 for age (included as a reference variable), cognitive ability and fluid intelligence, and sample sizes < 400 for all phenotypes except for inhibition measured by the flanker task (a measure that is known to have low reliability[14]).

Our results highlight, that the key determinant of sample size requirements is the true effect size of the brain–phenotype relationship, which subsumes the amount, quality, homogeneity and reliability of both brain and phenotypic measures, and the degree to which a particular brain measure is relevant to a particular phenotype. Effect sizes will probably vary widely across studies; for example, cortical thickness can also reliably predict 4 out of the 6 investigated phenotypes with $n < 500$, although with smaller effect sizes on average (functional connectivity, mean $r = 0.2$; cortical thickness, mean $r = 0.1$; Supplementary Fig. 2). Although our results were derived from a relatively high-quality dataset and used an algorithm expected to yield larger effect sizes than that of Marek et al.[4], they are in agreement with analytical calculations showing that BWAS that explain more than 1% of the phenotype's variance can be replicable with sample sizes below 1,000 (Supplementary Methods). For example, a model that explains $r^2 = 0.01$ (1% of variance) achieves 80% power in a prospective replication with $n = 801$, and $r^2 = 0.02$ achieves 80% power with $n = 399$ (ref. [15]).

These quantitative differences in required sample size could translate into large, qualitative differences in the types of neuroimaging studies considered viable in future efforts. There is a necessary trade-off between the innovativeness of a task, measure or method, and the extent to which it has been validated. Existing large-scale neuroimaging studies ($n > 1,000$) have selected well-validated tasks and imaging measures over new, exploratory ones, and few have attempted to characterize rare populations. Requiring sample sizes that are larger than necessary for the discovery of new effects could stifle innovation.

We agree with Marek et al.[4] that small-sample studies are important for understanding the brain bases of tasks and mental states[9–11], and for prototyping new tasks and measures. Furthermore, several current trends may further increase the viability of small-sample multivariate BWAS, including (1) new phenotypes, (2) feature-learning methods and algorithms with larger effect sizes[13], (3) models that target within-person variation in symptoms and behaviour to improve between-person predictions[2] and (4) hybrid strategies for improving prediction like meta-matching[16]. All of these have the potential to improve reliability and effect sizes, but whether they do remains to be seen.

Finally, as both Marek et al.[4] and our analyses show, very small effects will suffer from limited power, replicability and predictive utility even with sample sizes in the thousands (Fig. 1). We argue that the field should focus on discovering phenotypes and brain measures with large effect sizes. Efficient discovery entails casting a wide net in smaller studies using rigorous, unbiased methods and scaling up promising findings to larger samples[2]. There are substantial challenges ahead, including establishing broad generalizability across contexts, equity across subpopulations, and models with high neuroscientific validity and interpretability[17,18]. Addressing these challenges will require innovative new methods and measures.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-05745-x.

Reporting summary

Further information on experimental design is available in the Nature Portfolio Reporting Summary linked to this Article.

Data availability

Analysis is based on preprocessed data provided by the Human Connectome Project, WU-Minn Consortium (principal investigators: D. Van Essen and K. Ugurbil; 1U54MH091657) funded by the 16 NIH institutes and centres that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. All data used in the present study are available for download from the Human Connectome Project (www.humanconnectome.org). Users must agree to data use terms for the HCP before being allowed access to the data and ConnectomeDB; details are provided online (https://www.humanconnectome.org/study/hcp-young-adult/data-use-terms).

Code availability

All analysis code used in the current study is available at GitHub (release v.1.0; https://github.com/spisakt/BWAS_comment).

1. Button, K. S. et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
2. Woo, C.-W. W., Chang, L. J., Lindquist, M. A. & Wager, T. D. Building better biomarkers: Brain models in translational neuroimaging. *Nat. Neurosci.* **20**, 365–377 (2017).
3. Bzdok, D., Varoquaux, G. & Steyerberg, E. W. Prediction, not association, paves the road to precision medicine. *JAMA Psychiatry* **78**, 127–128 (2021).
4. Marek, S. et al. Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654–660 (2022).
5. Poldrack, R. A., Huckins, G. & Varoquaux, G. Establishment of best practices for evidence for prediction: a review. *JAMA Psychiatry* **77**, 534–540 (2020).
6. Genon, S., Eickhoff, S. B. & Kharabian, S. Linking interindividual variability in brain structure to behaviour. *Nat. Rev. Neurosci.* **23**, 307–318 (2022).
7. Rosenberg, M. D. & Finn, E. S. How to establish robust brain–behavior relationships without thousands of individuals. *Nat. Neurosci.* **25**, 835–837 (2022).
8. Van Essen, D. C. et al. The WU-Minn human connectome project: an overview. *Neuroimage*, **80**, 62–79 (2013).
9. Wager, T. D. et al. An fMRI-based neurologic signature of physical pain. *N. Engl. J. Med.* **368**, 1388–1397 (2013).
10. Han, X. et al. Effect sizes and test-retest reliability of the fMRI-based neurologic pain signature. *Neuroimage* **247**, 118844 (2022).
11. Zunhammer, M., Bingel, U. & Wager, T. D. Placebo effects on the neurologic pain signature. *JAMA Neurol.* **75**, 1321–1330 (2018).
12. Tian, Y. & Zalesky, A. Machine learning prediction of cognition from functional connectivity: are feature weights reliable? *Neuroimage* **245**, 118648 (2021).
13. Pervaiz, U., Vidaurre, D., Woolrich, M. W. & Smith, S. M. Optimising network modelling methods for fMRI. *Neuroimage* **211**, 116604 (2020).
14. Hedge, C., Powell, G. & Sumner, P. The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods* **50**, 1166–1186 (2018).

15. Killeen, P. R. Predict, control, and replicate to understand: how statistics can foster the fundamental goals of science. *Perspect. Behav. Sci.* **42**, 109–132 (2018).
16. He, T. et al. Meta-matching as a simple framework to translate phenotypic predictive models from big to small data. *Nat. Neurosci.* **25**, 795–804 (2022).
17. Wu, J. et al. Cross-cohort replicability and generalizability of connectivity-based psychometric prediction patterns. *Neuroimage* **262**, 119569 (2022).
18. Li, J. et al. Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Sci. Adv.* **8**, eabj1812 (2022).

Acknowledgements We thank S. Marek et al. for sharing the analysis code and for the discussions in relation to our commentary. The work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; projects 'TRR289 - Treatment Expectation', ID 422744262 and 'SFB1280 - Extinction Learning', ID 316803389), R01 MH076136 and R01 EB026549.

Author contributions Conception and data analysis: T.S. Manuscript writing and revision: T.S., U.B. and T.D.W.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-05745-x.

Correspondence and requests for materials should be addressed to Tamas Spisak.

Reprints and permissions information is available at http://www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

nature portfolio

Corresponding author(s): Tamas Spisak

Last updated by author(s): Dec 20, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Analysis is based on preprocessed data provided by the Human Connectome Project, WU-Minn Consortium (principal investigators: D. Van Essen and K. Ugurbil; 1U54MH091657) funded by the 16 NIH institutes and centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. All data used in the present study are available for download from the Human Connectome Project (www.humanconnectome.org). Users must agree to data use terms for the HCP before being allowed access to the data and ConnectomeDB; details are provided at https://www.humanconnectome.org/study/hcp-young-adult/data-use-terms. Preprocessed data was created software as described in: Glasser MF, Sotiropoulos SN, Wilson JA, Coalson TS, Fischl B, Andersson JL, Xu J, Jbabdi S, Webster M, Polimeni JR, Van Essen DC. The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage. 2013 Oct 15;80:105-24.
Data analysis	https://github.com/spisakt/BWAS_comment v1.0 Dependencies: python 3.10.8 numpy 1.23.5 pandas 1.5.2 scikit-learn 1.2.0 joblib 1.2.0 mlxtend 0.21.0 seaborn 0.12.1 matplotlib 3.6.2

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Analysis is based on preprocessed data provided by the Human Connectome Project, WU-Minn Consortium (principal investigators: D. Van Essen and K. Ugurbil; 1U54MH091657) funded by the 16 NIH institutes and centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. All data used in the present study are available for download from the Human Connectome Project (www.humanconnectome.org). Users must agree to data use terms for the HCP before being allowed access to the data and ConnectomeDB; details are provided at https://www.humanconnectome.org/study/hcp-young-adult/data-use-terms.
All derivative data, (including raw material for figures) is freely available at https://github.com/spisakt/BWAS_comment

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	The Human Connectome Project involved 656 females and 550 males. Findings apply to both males and females. Sex and gender differences are out of the scope of the Matters Arising and were not considered in the analysis. For more information, refer to van Essen et al., 2013.
Population characteristics	Refer to van Essen et al., 2013.
Recruitment	As described in van Essen et al., 2013.
Ethics oversight	As described in van Essen et al., 2013.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Quantitative analyses of the replicability of functional connectivity-based brain-wise association studies.
Research sample	We re-analyzed one of the studies involved in the original publication by Marek, Tervo-Clemmens et al., including open access-data from the Human Connectome Project (van Essen et al., 2013), based on a non-representative sample of young adults (656 females, 550 males, mean+-sd age: 28.9+-3.57).
Sampling strategy	We re-analyzed one of the studies involved in the original publication by Marek, Tervo-Clemmens et al., including open access-data from the Human Connectome Project. For more information, please refer to van Essen et al., 2013 ands Marek, Tervo-Clemmens et al, 2022.
Data collection	See van Essen et al., 2013 ands Marek, Tervo-Clemmens et al, 2022 for details.
Timing	See van Essen et al., 2013 ands Marek, Tervo-Clemmens et al, 2022 for details.
Data exclusions	We excluded participants with no MRI images available, and for each analysis, participants with missing data about the target phenotype.
Non-participation	203 out of the 1206 HCP participants didn't have MRI data. Exclusion die to missing phenotype data varied across analyses (less than 20 in all cases)
Randomization	N/A (observational study)

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Magnetic resonance imaging

Experimental design

Design type: resting state fMRI, anatomical MRI

Design specifications: 4, 15min, eyes open (see van Essen et al., 2013)

Behavioral performance measures: As described in van Essen et al., 2013.

Acquisition

Imaging type(s): functional, anatomical

Field strength: 3

Sequence & imaging parameters: As described in van Essen et al., 2013.

Area of acquisition: brain

Diffusion MRI: ☐ Used ☒ Not used

Preprocessing

Preprocessing software: As described in:
Glasser MF, Sotiropoulos SN, Wilson JA, Coalson TS, Fischl B, Andersson JL, Xu J, Jbabdi S, Webster M, Polimeni JR, Van Essen DC. The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage. 2013 Oct 15;80:105-24.

Normalization: As described in: van Essen et al., 2013.

Normalization template: As described in: van Essen et al., 2013.

Noise and artifact removal: As described in: van Essen et al., 2013.

Volume censoring: As described in: van Essen et al., 2013.

Statistical modeling & inference

Model type and settings: Multivariate predictive modelling, PCA+SVR and Ridge, cross-validation (see Supplementary Methods for details).

Effect(s) tested: Predictive performance (Pearson's r)

Specify type of analysis: ☒ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016): N/A (predictive modelling)

Correction: N/A (predictive modelling)

Models & analysis

n/a	Involved in the study
☐	☒ Functional and/or effective connectivity
☒	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

partial correlation

Multivariate modeling and predictive analysis

Targets: age, cognitive ability, episodic memory, fluid intelligence, cognitive flexibility, inhibition. Model: PCA +SVR (reproduction of Marek et al.'s model), Ridge regression (with the default hyperparameter value 1). The PCA+SVR model involved dimensionality reduction.
Features partial correlation values across 100 ICA-based regions, training metric: mean squared error, evaluation metric, mean squared error and Pearson correlation.

Reply to: Multivariate BWAS can be replicable with moderate sample sizes

https://doi.org/10.1038/s41586-023-05746-w
Published online: 8 March 2023
Open access
Check for updates

Brenden Tervo-Clemmens[1,22]✉, Scott Marek[2,3,22]✉, Roselyne J. Chauvin[4], Andrew N. Van[4,5], Benjamin P. Kay[4], Timothy O. Laumann[3], Wesley K. Thompson[6], Thomas E. Nichols[7], B. T. Thomas Yeo[8,9,10,11,12,13], Deanna M. Barch[3,14], Beatriz Luna[15,16], Damien A. Fair[17,18,19,23]✉ & Nico U. F. Dosenbach[2,4,5,14,20,21,23]✉

REPLYING TO: T. Spisak et al. *Nature* https://doi.org/10.1038/s41586-023-05745-x (2023)

In our previous study[1], we documented the effect of sample size on the reproducibility of brain-wide association studies (BWAS) that aim to cross-sectionally relate individual differences in human brain structure (cortical thickness) or function (resting-state functional connectivity (RSFC)) to cognitive or mental health phenotypes. Applying univariate and multivariate methods (for example, support vector regression (SVR)) to three large-scale neuroimaging datasets (total $n \approx 50{,}000$), we found that overall BWAS reproducibility was low for $n < 1{,}000$, due to smaller than expected effect sizes. When samples and true effects are small, sampling variability, and/or overfitting can generate 'statistically significant' associations that are likely to be reported due to publication bias, but are not reproducible[2–5], and we therefore suggested that BWAS should build on recent precedents[6,7] and continue to aim for samples in the thousands. In the accompanying Comment, Spisak et al.[8] agree that larger BWAS are better[5,9], but argue that "multivariate BWAS effects in high-quality datasets can be replicable with substantially smaller sample sizes in some cases" ($n = 75$–500); this suggestion is made on the basis of analyses of a selected subset of multivariate cognition/RSFC associations with larger effect sizes, using their preferred method (ridge regression with partial correlations) in a demographically more homogeneous, single-site/scanner sample (Human Connectome Project (HCP), $n = 1{,}200$, aged 22–35 years).

There is no disagreement that a minority of BWAS effects can replicate in smaller samples, as shown with our original methods[1]). Using the exact methodology (including cross-validation) and code of Spisak et al.[8] to repeat 64 multivariate BWAS in the 21-site, larger and more diverse Adolescent Brain Cognitive Development Study (ABCD, $n = 11{,}874$, aged 9–11 years), we found that 31% replicated at $n = 1{,}000$, dropping to 14% at $n = 500$ and none at $n = 75$. Contrary to the claims of Spisak et al.[8], replication failure was the outcome in most cases when applied to this larger, more diverse dataset. Basing general BWAS sample size recommendations on the largest effects has at least two fundamental flaws: (1) failing to detect other true effects (for example, reducing the sample size from $n = 1{,}000$ to $n = 500$ leads to a 55% false-negative rate), therefore restricting BWAS scope, and (2) inflation of reported effects[3,10–12]. Thus, regardless of the method, associations based on small samples can remain distorted and lack generalizability until confirmed in large, diverse, independent samples.

We always test for BWAS replication with null models (using permutation tests) of out-of-sample estimates to ensure that our reported reproducibility is unaffected by in-sample overfitting. Nonetheless, Spisak et al.[8] argue against plotting inflated in-sample estimates[1,10] on the y axis, and out-of-sample values on the x axis, as we did (Fig. 1a). Instead, they propose plotting cross-validated associations from an initial, discovery sample (Fig. 1b (y axis)) against split-half out-of-sample associations (x axis). However, cross-validation—just like split-half validation—estimates out-of-sample, and not in-sample, effect sizes[13]. The in-sample associations[1,10] for the method of Spisak et al.[8] (Fig. 1b), that is, from data in the sample used to develop the model, show the same degree of overfitting (Fig. 1a versus Fig. 1b). The plot of Spisak et al.[8] (Fig. 1c) simply adds an additional out-of-sample test (cross-validation before split half), and therefore demonstrates the close correspondence between two different methods for out-of-sample effect estimation[14]. Analogously, we can replace the cross-validation step in the code of Spisak et al.[8] with split-half validation (our original out-of-sample test), obtaining split-half effects in the first half of the sample, and then comparing them to the split-half estimates from the full sample (Fig. 1d). The strong correspondences between cross-validation followed by split-half (Spisak et al. method[8]; Fig. 1c) and repeated split-half validation (Fig. 1d) are guaranteed by plotting out-of-sample estimates (from the same dataset) against one another. Here, plotting cross-validated discovery sample estimates on the y axis (Fig. 1c,d) provides no additional information beyond the x axis out-of-sample values. The critically important out-of-sample predictions, required for reporting multivariate results[1], generated using the method of Spisak et al.[8] and our method are nearly identical (Fig. 1e).

As Spisak et al.[8] highlight, cross-validation of some type is considered to be standard practice[10], and yet the distribution of out-of-sample

[1]Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [2]Department of Radiology, Washington University School of Medicine, St Louis, MO, USA. [3]Department of Psychiatry, Washington University School of Medicine, St Louis, MO, USA. [4]Department of Neurology, Washington University School of Medicine, St Louis, MO, USA. [5]Department of Biomedical Engineering, Washington University in St Louis, St Louis, MO, USA. [6]Division of Biostatistics, University of California San Diego, La Jolla, CA, USA. [7]Oxford Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of Oxford, Oxford, UK. [8]Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore. [9]Centre for Sleep and Cognition, National University of Singapore, Singapore, Singapore. [10]Centre for Translational MR Research, National University of Singapore, Singapore, Singapore. [11]N.1 Institute for Health, Institute for Digital Medicine, National University of Singapore, Singapore, Singapore. [12]Integrative Sciences and Engineering Programme, National University of Singapore, Singapore, Singapore. [13]Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA, USA. [14]Department of Psychological and Brain Sciences, Washington University in St Louis, St Louis, MO, USA. [15]Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA. [16]Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA. [17]Masonic Institute for the Developing Brain, University of Minnesota Medical School, Minneapolis, MN, USA. [18]Department of Pediatrics, University of Minnesota Medical School, Minneapolis, MN, USA. [19]Institute of Child Development, University of Minnesota Medical School, Minneapolis, MN, USA. [20]Program in Occupational Therapy, Washington University School of Medicine, St Louis, MO, USA. [21]Department of Pediatrics, Washington University School of Medicine, St Louis, MO, USA. [22]These authors contributed equally: Brenden Tervo-Clemmens, Scott Marek. [23]These authors jointly supervised this work: Damien A. Fair, Nico U. F. Dosenbach. ✉e-mail: btervo-clemmens@mgh.harvard.edu; smarek@wustl.edu; faird@umn.edu; ndosenbach@wustl.edu

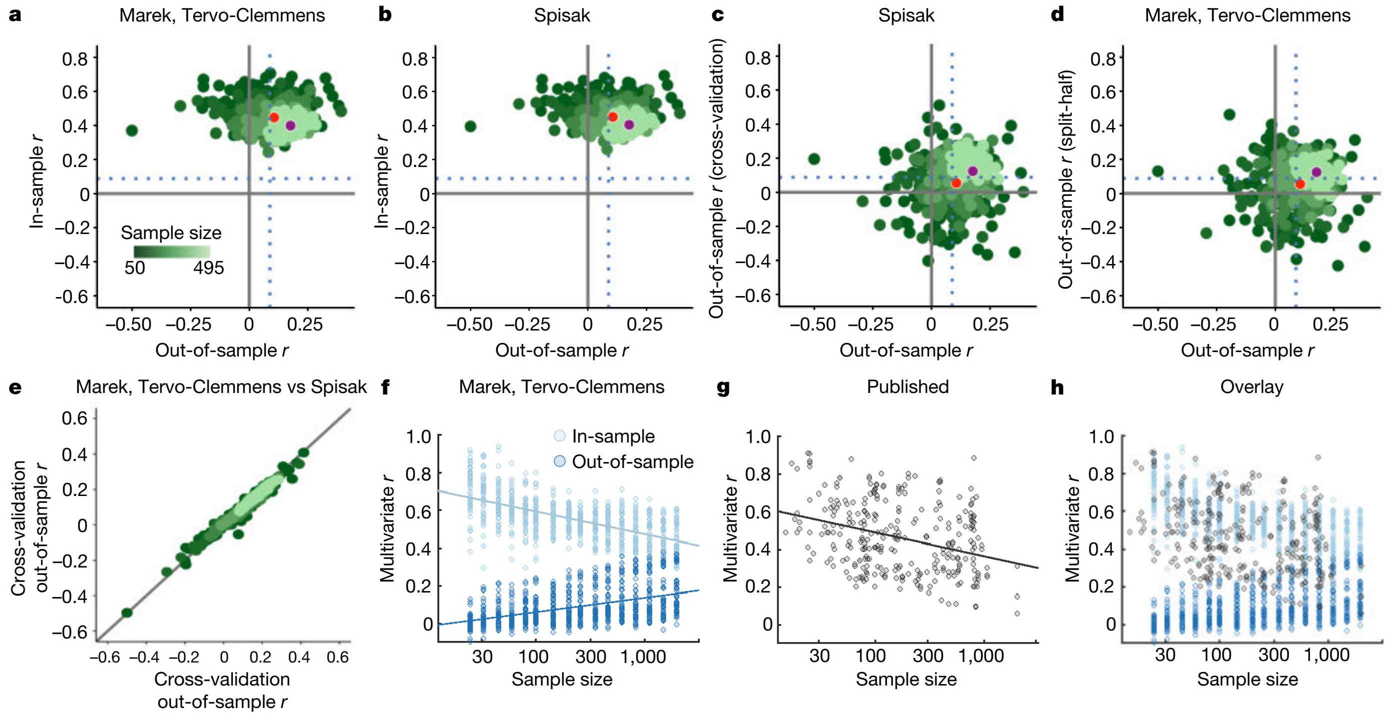


Fig. 1 | In-sample versus out-of-sample effect estimates in multivariate BWAS. **a**–**e**, Methods comparison between our previous study[1] (split-half) and Spisak et al.[8] (cross-validation followed by split-half). 'Marek, Tervo-Clemmens' and 'Spisak' refer to the methodolgies described in ref. [1] and ref. [8], respectively. For **a**–**e**, HCP 1200 Release (full correlation) data were used to predict age-adjusted total cognitive ability. Analysis code and visualizations (*x*,*y* scaling; colours) are the same as in Spisak et al.[8]. The *x* axes in **a**–**e** always display the split-half out-of-sample effect estimates from the second (replication) half of the data (correlation between true scores and predicted scores; as in Spisak et al.[8] and in our previous study[1]; Supplementary Methods). **a**, In-sample (training correlation; *y* axis) as a function of out-of-sample associations (plot convention in our previous study[1]). **b**, Matched comparison of the true in-sample association (training correlations, mean across folds; *y* axis) in the method proposed by Spisak et al.[8]. **c**, The proposed correction by Spisak et al.[8] that inserts an additional cross-validation step to evaluate the first half of data, which by definition makes this an out-of-sample association (*y* axis). **d**, Replacing the cross-validation step from Spisak et al.[8] with a split-half validation provides a different (compared with **c**) out-of-sample association of the first half of the total data (that is, each of the first stage split halves is one-quarter of the total data; *y* axis). The appropriate and direct comparison of in-sample associations between Spisak et al.[8] and our previous study[1] is comparing **b** to **a**, rather than **c** to **a**. The Spisak et al. method[8] (cross-validation followed by split-half validation) does not reduce in-sample overfitting (**b**) but, instead, adds an additional out-of-sample evaluation (**c**), which is nearly identical to split-half validation twice in a row (**d**), and makes it clear why the out-of-sample performance of these two methods is likewise nearly identical. **e**, Correspondence between out-of-sample associations (to the left-out half) from the additional cross-validation step proposed by Spisak et al.[8] (mean across folds; *y* axis) and the original split-half validation from our previous study[1] (*x* axis). The identity line is shown in black. **f**, In-sample (*r*; light blue) and out-of-sample (*r*; dark blue) associations as a function of sample size. Data are from figure 4a–d of ref. [1]. **g**, Published literature review of multivariate *r* (*y* axis) as a function of sample size (data from ref. [15]) displayed with permission. For **f** and **g**, best fit lines are displayed in $\log_{10}$ space. **h**, Overlap of **f** and **g**.

associations (Fig. 1f (dark blue)) does not match published multivariate BWAS results (Fig. 1g), which have largely ranged from $r = 0.25$ to 0.9, decreasing with increasing sample size[10,15,16]. Instead, published effects more closely follow the distribution of in-sample associations (Fig. 1h). This observation suggests that, in addition to small samples, structural problems in academic research (for example, non-representative samples, publication bias, misuse of cross-validation and unintended overfitting) have contributed to the publication of inflated effects[12,17,18]. A recent biomarker challenge[5] showed that cross-validation results continued to improve with the amount of time researchers spent with the data, and the models with the best cross-validation results performed worse on never-seen held-back data. Thus, cross-validation alone has proven to be insufficient and must be combined with the increased generalizability of large, diverse datasets and independent out-of-sample evaluation in new, never before seen data[5,10].

The use of additional cross-validation in the discovery sample by Spisak et al.[8] does not affect out-of-sample prediction accuracies (Fig. 1e). However, by using partial correlations and ridge regression on HCP data, they were able to generate higher out-of-sample prediction accuracies than our original results in ABCD (Fig. 2a). The five variables they selected are strongly correlated[19] cognitive measures from the NIH Toolbox (mean $r = 0.37$; compare with the correlation strength for height versus weight $r = 0.44$)[20] and age (not a complex behavioural phenotype), unrepresentative of BWAS as a whole (Fig. 2b (colour versus grey lines)). As the HCP is the relatively smallest and most homogeneous dataset, we applied the exact method and code of Spisak et al.[8] to the ABCD data (Fig. 2c and Supplementary Table 2). At $n = 1{,}000$ (training; $n = 2{,}000$ total), only 31% of BWAS (44% RSFC, 19% cortical thickness) were replicable (Fig. 2d; defined as in Spisak et al.[8]; Supplementary Information). Expanding BWAS scope beyond broad cognitive abilities towards complex mental health outcomes therefore requires $n > 1{,}000$ (Fig. 2b–d). The absolute largest BWAS (cognitive ability: RSFC, green) reached replicability only using $n = 400$ ($n = 200$ train; $n = 200$ test) with an approximate 40% decrease in out-of-sample prediction accuracies from HCP to ABCD (Fig. 2e (lighter green, left versus right)). The methods of Spisak et al.[8] and our previous study[1] returned equivalent out-of-sample reproducibility for this BWAS (cognitive ability: RSFC) in the larger, more diverse ABCD data (Fig. 2e (right, dark versus light green)). Thus, the smaller sample sizes (Fig. 2b,c) that are required for out-of-sample reproducibility (Fig. 2e) reported by Spisak et al.[8] in the HCP data did not generalize to the larger ABCD dataset. See also our previous study[1] for a broader discussion of convergent evidence across HCP and ABCD datasets.

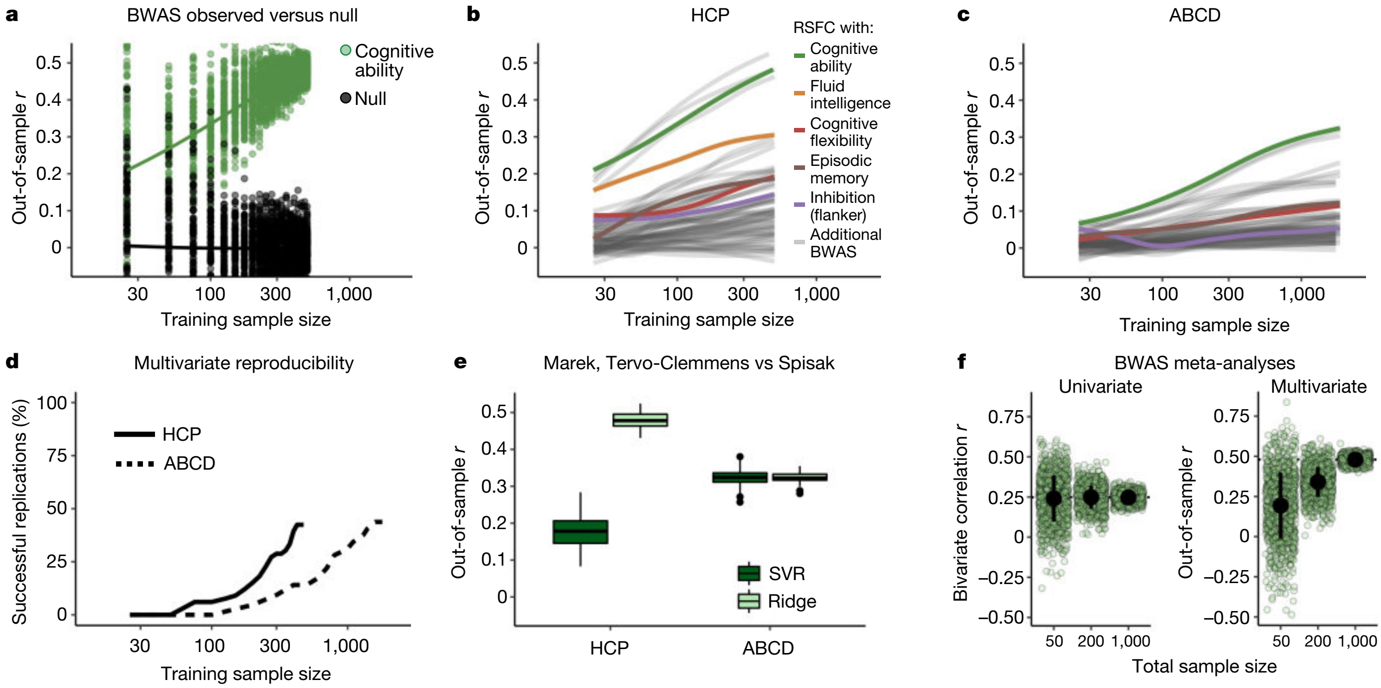


Fig. 2 | BWAS reproducibility, scope and prediction accuracy using the method of Spisak et al. a, Example bootstrapped BWAS of total cognitive ability (green) and null distribution (black) (*y* axis), as a function of sample size (*x* axis) from the suggested method of Spisak et al.[8] (RSFC by partial correlation; prediction by ridge regression) in the HCP dataset ($n = 1{,}200$, 1 site, 1 scanner, 60 min RSFC/participant, 76% white). Sample sizes were $\log_{10}$-transformed for visualization. **b**, Out-of-sample correlation (between true scores and predicted scores) from ridge regression (*y* axis; code from Spisak et al.[8]) as a function of training sample size (*x* axis, $\log_{10}$ scaling) for 33 cognitive and mental health phenotypes (Supplementary Information) in the HCP dataset. Each line displays a smoothed fit estimate (through penalized splines in general additive models) for a brain (RSFC (partial correlations, as proposed by Spisak et al.[8]), cortical thickness) phenotype pair (66 total) that has 100 bootstrapped iterations from sample sizes of 25 to 500 (inclusive) in increments of 25 (20 total bins). Sample sizes were $\log_{10}$-transformed (for visualization) before general additive model fitting. **c**, The same as in **b**, but in the ABCD dataset ($n = 11{,}874$, 21 sites, 3 scanner manufacturers, 20 min RSFC/participant, 56% white) using 32 cognitive and mental health phenotypes at sample sizes of 25, 50, 75 and from 100 to 1,900 (inclusive) in increments of 100 (22 total bins). **d**, The percentage of brain–phenotype pairs (BWAS) from **b** and **c** with significant replication on the basis of the method of Spisak et al.[8] (Supplementary Information). **e**, Comparison of our original method in our previous study[1] and the method proposed by Spisak et al.[8] at the full split-half sample size of HCP (left) and ABCD (right). Out-of-sample correlations (RSFC with total cognitive ability, *y* axis) for the method used in our previous study[1] (dark green; RSFC by correlation, PCA, SVR) and by Spisak et al.[8] (light green; RSFC by partial correlation, ridge regression). Repeating the method proposed by Spisak et al.[8] in ABCD (right) and comparing this to the method used in our previous study[1] results in a very similar out-of-sample *r*. **f**, Simulated individual studies (light green circles; $n = 1{,}000$ per sample size) and meta-analytic estimates (black dot, ±1 s.d.) using the method of Spisak et al.[8] (partial correlations in the HCP dataset) for the largest univariate association (left; *y* axis, bivariate correlation) and multivariate association (right; *y* axis, out-of-sample correlation) for total cognitive ability versus RSFC, as a function of total sample size (*x* axis; bivariate correlation for sample sizes of 50, 200 and 1,000, and multivariate sum of train and test samples, each 25, 100 and 500). For univariate approaches, studies of any sample size, when appropriately aggregated to a large total sample size, can correctly estimate the true effect size. However, for multivariate approaches, even when aggregating across 1,000 independent studies, studies with a small sample size produce prediction accuracies that are downwardly biased relative to large sample studies, highlighting the need for large samples in multivariate analyses.

Notably, the objections of Spisak et al.[8] raise additional reasons to stop the use of smaller samples in BWAS that were not highlighted in our original article. Multivariate BWAS prediction accuracies—absent overfitting—are systematically suppressed in smaller samples[5,9,21], as prediction accuracy scales with increasing sample size[1,9]. Thus, the claim that "cross-validated discovery effect-size estimates are unbiased" does not account for out-of-dataset generalizability and downward bias. In principle, if unintended overfitting and publication bias could be fully eliminated, meta-analyses of small-sample univariate BWAS would return the correct association strengths (Fig. 2f (left)). However, meta-analyses of small multivariate BWAS would always be downwardly biased (Fig. 2f (right)). If we are interested in maximizing prediction accuracy, essential for clinical implementation of BWAS[22], large samples and advancements in imaging and phenotypic measurements[1] are necessary.

Repeatedly subsampling the same dataset, as Spisak et al.[8] and we have done, overestimates reproducibility compared with testing on a truly new, diverse dataset. Just as in genomics[23], BWAS generalization failures have been highlighted[5,24]. For example, BWAS models trained on white Americans transferred poorly to African Americans and vice versa (within dataset)[24]. Historically, BWAS samples have lacked diversity, neglecting marginalized and under-represented minorities[25]. Large studies with more diverse samples and data aggregation efforts can improve BWAS generalizability and reduce scientific biases contributing to massive health inequities[26,27].

Spisak et al.[8] worry that "[r]equiring sample sizes that are larger than necessary for the discovery of new effects could stifle innovation". We appreciate the concern that rarer populations may never be investigated with BWAS. Yet, there are many non-BWAS brain–behaviour study designs (fMRI ≠ BWAS) focused on within-patient effects, repeated-sampling and signal-to-noise-ratio improvements that have proven fruitful down to $n = 1$ (ref. [28]). By contrast, the strength of multivariate BWAS lies in leveraging large cross-sectional samples to investigate population-level questions. Sample size requirements should be based on expected effect sizes and real-world impact, and not resource availability. Through large-scale collaboration and clear standards on data sharing, GWAS has reached sample sizes in the millions[29–31], pushing genomics towards new horizons. Similarly, BWAS

analyses of the future will not be limited to statistical replication of the same few strongest effects in small homogeneous populations, but also have broad scope, maximum prediction accuracy and excellent generalizability.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-05746-w.

Reporting summary

Further information on experimental design is available in the Nature Portfolio Reporting Summary linked to this Article.

Data availability

Participant-level data from all datasets (ABCD and HCP) are openly available pursuant to individual, consortium-level data access rules. The ABCD data repository grows and changes over time (https://nda.nih.gov/abcd). The ABCD data used in this report came from ABCD collection 3165 and the Annual Release 2.0 (https://doi.org/10.15154/1503209). Data were provided, in part, by the HCP, WU-Minn Consortium (principal investigators: D. Van Essen and K. Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. Some data used in the present study are available for download from the HCP (www.humanconnectome.org). Users must agree to data use terms for the HCP before being allowed access to the data and ConnectomeDB, details are provided online (https://www.humanconnectome.org/study/hcp-young-adult/data-use-terms).

Code availability

Manuscript analysis code specific to this study is available at GitHub (https://gitlab.com/DosenbachGreene/bwas_response). Code for processing ABCD data is provided at GitHub (https://github.com/DCAN-Labs/abcd-hcp-pipeline). MRI data analysis code is provided at GitHub (https://github.com/ABCD-STUDY/nda-abcd-collection-3165). FIRMM software is available online (https://firmm.readthedocs.io/en/latest/release_notes/). The ABCD Study used v.3.0.14.

1. Marek, S. et al. Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654–660 (2022).
2. Schönbrodt, F. D. & Perugini, M. At what sample size do correlations stabilize? *J. Res. Pers.* **47**, 609–612 (2013).
3. Button, K. S. et al. Confidence and precision increase with high statistical power. *Nat. Rev. Neurosci.* **14**, 585–586 (2013).
4. Varoquaux, G. Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage* **180**, 68–77 (2018).
5. Traut, N. et al. Insights from an autism imaging biomarker challenge: promises and threats to biomarker discovery. *Neuroimage* **255**, 119171 (2022).
6. Casey, B. J. et al. The Adolescent Brain Cognitive Development (ABCD) study: imaging acquisition across 21 sites. *Dev. Cogn. Neurosci.* **32**, 43–54 (2018).
7. Littlejohns, T. J. et al. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nat. Commun.* **11**, 2624 (2020).
8. Spisak, T., Bingel, U. & Wager, T. D. Multivariate BWAS can be replicable with moderate sample sizes. *Nature* https://doi.org/10.1038/s41586-023-05745-x (2023).
9. Schulz, M.-A., Bzdok, D., Haufe, S., Haynes, J.-D. & Ritter, K. Performance reserves in brain-imaging-based phenotype prediction. Preprint at https://doi.org/10.1101/2022.02.23.481601 (2022).
10. Poldrack, R. A., Huckins, G. & Varoquaux, G. Establishment of best practices for evidence for prediction: a review. *JAMA Psychiatry* **77**, 534–540 (2020).
11. Poldrack, R. A. The costs of reproducibility. *Neuron* **101**, 11–14 (2019).
12. Button, K. S. et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
13. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proc. IJCAI 95* (ed. Mellish, C. S.) 1137–1143 (Morgan Kaufman, 1995).
14. Scheinost, D. et al. Ten simple rules for predictive modeling of individual differences in neuroimaging. *Neuroimage* **193**, 35–45 (2019).
15. Sui, J., Jiang, R., Bustillo, J. & Calhoun, V. Neuroimaging-based individualized prediction of cognition and behavior for mental disorders and health: methods and promises. *Biol. Psychiatry* **88**, 818–828 (2020).
16. Woo, C.-W., Chang, L. J., Lindquist, M. A. & Wager, T. D. Building better biomarkers: brain models in translational neuroimaging. *Nat. Neurosci.* **20**, 365–377 (2017).
17. Ioannidis, J. P. A. Why most discovered true associations are inflated. *Epidemiology* **19**, 640–648 (2008).
18. Pulini, A. A., Kerr, W. T., Loo, S. K. & Lenartowicz, A. Classification accuracy of neuroimaging biomarkers in attention-deficit/hyperactivity disorder: effects of sample size and circular analysis. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **4**, 108–120 (2019).
19. Funder, D. C. & Ozer, D. J. Evaluating effect size in psychological research: sense and nonsense. *Adv. Methods Pract. Psychol. Sci.* **2**, 156–168 (2019).
20. Meyer, G. J. et al. Psychological testing and psychological assessment: a review of evidence and issues. *Am. Psychol.* **56**, 128–165 (2001).
21. He, T. et al. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *Neuroimage* **206**, 116276 (2020).
22. Leptak, C. et al. What evidence do we need for biomarker qualification? *Sci. Transl. Med.* **9**, eaal4599 (2017).
23. Weissbrod, O. et al. Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nat. Genet.* **54**, 450–458 (2022).
24. Li, J. et al. Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Sci. Adv.* **8**, eabj1812 (2022).
25. Henrich, J., Heine, S. J. & Norenzayan, A. The weirdest people in the world? *Behav. Brain Sci.* **33**, 61–83 (2010).
26. Bailey, Z. D. et al. Structural racism and health inequities in the USA: evidence and interventions. *Lancet* **389**, 1453–1463 (2017).
27. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
28. Gratton, C., Nelson, S. M. & Gordon, E. M. Brain-behavior correlations: two paths toward reliability. *Neuron* **110**, 1446–1449 (2022).
29. Levey, D. F et al. Bi-ancestral depression GWAS in the Million Veteran Program and meta-analysis in >1.2 million individuals highlight new therapeutic directions. *Nature* **24**, 954–963 (2021).
30. Muggleton, N. et al. The association between gambling and financial, social and health outcomes in big financial data. *Nat. Hum. Behav.* **5**, 319–326 (2021).
31. Yengo, L. et al. A saturated map of common genetic variants associated with human height. *Nature* **610**, 704–712 (2022).

Acknowledgements Data used in the preparation of this Article were, in part, obtained from the Adolescent Brain Cognitive Development (ABCD) Study (https://abcdstudy.org), held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children aged 9–10 years and follow them over 10 years into early adulthood. The ABCD Study is supported by the National Institutes of Health and additional federal partners under award numbers U01DA041022, U01DA041028, U01DA041048, U01DA041089, U01DA041106, U01DA041117, U01DA041120, U01DA041134, U01DA041148, U01DA041156, U01DA041174, U24DA041123, U24DA041147, U01DA041093 and U01DA041025. A full list of supporters is available online (https://abcdstudy.org/federal-partners.html). A listing of participating sites and a complete listing of the study investigators is available online (https://abcdstudy.org/scientists/workgroups/). ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in the analysis or the writing of this report. This Article reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators. Data were provided, in part, by the HCP, WU-Minn Consortium (U54 MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. This work used the storage and computational resources provided by the Masonic Institute for the Developing Brain (MIDB), the Neuroimaging Genomics Data Resource (NGDR) and the Minnesota Supercomputing Institute (MSI). The NGDR is supported by the University of Minnesota Informatics Institute through the MnDRIVE initiative in coordination with the College of Liberal Arts, Medical School and College of Education and Human Development at the University of Minnesota. This work used the storage and computational resources provided by the Daenerys Neuroimaging Community Computing Resource (NCCR). The Daenerys NCCR is supported by the McDonnell Center for Systems Neuroscience at Washington University, the Intellectual and Developmental Disabilities Research Center (IDDRC; P50 HD103525) at Washington University School of Medicine and the Institute of Clinical and Translational Sciences (ICTS; UL1 TR002345) at Washington University School of Medicine. This work was supported by NIH grants MH121518 (to S.M.), NS090978 (to B.P.K.), MH129616 (to T.O.L.), 1RF1MH120025-01A1 (to W.K.T), MH080243 (to B.L.), MH067924 (to B.L.), DA041148 (to D.A.F.), DA04112 (to D.A.F.), MH115357 (to D.A.F.), MH096773 (to D.A.F. and N.U.F.D.), MH122066 (to D.A.F. and N.U.F.D.), MH121276 (to D.A.F. and N.U.F.D.), MH124567 (to D.A.F. and N.U.F.D.), NS088590 (to N.U.F.D.), and the Andrew Mellon Predoctoral Fellowship (to B.T.-C.), the Staunton Farm Foundation (to B.L.), the Lynne and Andrew Redleaf Foundation (to D.A.F.) and the Kiwanis Neuroscience Research Foundation (to N.U.F.D.).

Author contributions Conception: B.T.-C., S.M., D.A.F. and N.U.F.D. Design: B.T.-C., S.M., R.J.C., D.A.F. and N.U.F.D. Data acquisition, analysis and interpretation: B.T.-C., S.M., R.J.C., A.V.N., B.P.K., W.K.T., T.E.N., B.T.T.Y., D.A.F. and N.U.F.D. Manuscript writing and revising: B.T.-C., S.M., R.J.C., A.V.N., B.P.K., T.O.L., W.K.T., T.E.N., B.T.T.Y., D.M.B., B.L., D.A.F. and N.U.F.D. We note that the reply author list differs from the original paper in number and in order to accurately reflect its more focused scope compared with the original work.

Competing interests D.A.F. and N.U.F.D. have a financial interest in Turing Medical and may financially benefit if the company is successful in marketing FIRMM motion monitoring software products. A.N.V., D.A.F. and N.U.F.D. may receive royalty income based on FIRMM technology developed at Washington University School of Medicine and Oregon Health and Sciences University and licensed to Turing Medical. D.A.F. and N.U.F.D. are co-founders of Turing Medical. These potential conflicts of interest have been reviewed and are managed by Washington University School of Medicine, Oregon Health and Sciences University and the University of Minnesota.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-05746-w.
Correspondence and requests for materials should be addressed to Brenden Tervo-Clemmens, Scott Marek, Damien A. Fair or Nico U. F. Dosenbach.
Reprints and permissions information is available at http://www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

nature portfolio

Corresponding author(s): Scott Marek, Brenden Tervo-Clemmens, Nico Dosenbach

Last updated by author(s): Dec 9, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection. Neuroimaging and behavioral data were from existing, open-source datasets (ABCD, UKB, HCP) whose acquisition's are presented in detail in previous work. The ABCD Study data were collected between 2016-2018. The HCP data were collected between 2010-2016.
Data analysis	MRI data analysis code can be found here: https://github.com/ABCD-STUDY/nda-abcd-collection-3165 ABCD and UKB MRI data processing code can be found here https://github.com/DCAN-Labs/abcd-hcp-pipeline Manuscript analysis code can be found here https://gitlab.com/DosenbachGreene/bwas_response FIRMM software: https://firmm.readthedocs.io/en/latest/release_notes/. ABCD uses version 3.0.14.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Participant level data from all datasets (ABCD & HCP) is openly available pursuant to individual, consortia-level data access rules. The ABCD data repository grows and changes over time. The ABCD data used in this report came from ABCD collection 3165 and the Annual Release 2.0, DOI 10.15154/1503209.

Data were provided, in part, by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. Some data used in the present study are available for download from the Human Connectome Project (www.humanconnectome.org). Users must agree to data use terms for the HCP before being allowed access to the data and ConnectomeDB, details are provided at https://www.humanconnectome.org/study/hcp-young-adult/data-use-terms.
No new data were collected for this manuscript. Across the ABCD, and HCP we downloaded data between 01/2019 - 10/2021. We did not use any specific software for downloading the data. For details on data collection in ABCD (baseline data), see Casey et al., 2018; in HCP (1200 release) see Van Essen et al., 2013).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Quantitative analyses of the magnitude and reproducibility of cross-sectional associations between neuroimaging measures and psychological/psychiatric phenotypes.
Research sample	Our main focus was to replicate work from Spisak et al in both the dataset used in their paper (HCP) and to further test generalizability of their models. Therefore, we also tested their models on the larger ABCD dataset.
Sampling strategy	All samples were recruited from the community (ABCD & HCP from the USA). Individual samples (ABCD, HCP) used unique sample size calculations and sampling strategies which are discussed in prior work with these open source datasets (Casey et al., 2018, Van Essen et al., 2013, respectively).
Data collection	All data were from existing data repositories and were downloaded between 01/2019 - 10/2021. Data used in the manuscript were from existing large consortia datasets (ABCD: see Casey et al., 2018 & Barch et al., 2018; HCP: We used data from the 1200 subjects data release (van Essen et al., 2013). Because we did not personally collect any of the data used in this manuscript, all data were from existing data repositories and researchers were therefore not blind to the source of the data.
Timing	ABCD: see Casey et al., 2018 HCP: see van Essen et al., 2013
Data exclusions	In ABCD, we used strict inclusion criteria with regard to head motion. Specifically, inclusion criteria for the current project consisted of at least 600 frames (8 minutes) of low-motion (filtered FD<0.08) resting state functional connectivity data. Our final dataset consisted of data from a total of N=3,928 youth across the discovery (N=1,964) and replication (N=1,964) sets. The final discovery and replication sets did not differ in mean FD (ΔM=0.002 , t=0.60, p=0.55) or total frames included (ΔM=6.4 , t =0.94, p=0.35). The subject lists for ARMS samples and our associated matrices will be released in the ABCD-BIDS Community Collection (ABCD collection 3165) for community use. For HCP data, we used similar data quantity inclusion, as well as an FD < 0.20 (unfiltered FD). This resulted in the inclusion of N=900 individuals (N=877 across all NIH Toolbox subscales).
Non-participation	N/A
Randomization	All three samples were observational studies and no randomization was used.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics: See above.

Recruitment: See above.

Ethics oversight: The ABCD Study obtained centralized institutional review board approval from the University of California, San Diego, and each of the 21 study sites obtained local institutional review board approval. Ethical regulations were followed during data collection and analysis. Parents or caregivers provided written informed consent, and children gave written assent.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Magnetic resonance imaging

Experimental design

Design type: resting-state fMRI, task-based fMRI; structural (cortical thickness) MRI

Design specifications: ABCD resting state: 4, 5 min runs, eyes open
HCP resting state: 4, 15 min runs, eyes open

Behavioral performance measures: Primary analyses use cognitive assessments from the NIH Toolbox and psychopathology assessment Child Behavior Checklist (see manuscript for individual subscales, total of 41) included in standard data releases and discussed in detail perviously (Barch et al., 2018)

Acquisition

Imaging type(s): Resting-state fMRI, task-fMRI, structural (cortical thickness) MRI

Field strength: 3 Tesla

Sequence & imaging parameters: Primary analyses use open-source distributed fMRI and MR data that adhere to consortia guidelines (see Casey et al., 2018 and Van Essen et al., 2013, for ABCD and HCP, respectively).

Area of acquisition: Whole brain

Diffusion MRI: ☐ Used ☒ Not used

Preprocessing

Preprocessing software: Preprocessing of ABCD was done using a suite of tools. All code can be found here: https://github.com/ABCD-STUDY/nda-abcd-collection-3165. Individual datasets (ABCD, HCP) and individual study sites (e.g., ABCD site 1 versus site 2) used unique sequence and imaging parameters which are discussed in prior work introducing these open-source datasets.

Normalization: 1) PreFreesurfer normalizes anatomical data. This normalization entails brain extraction, denoising, and then bias field correction on anatomical T1 and/or T2 weighted data. The ABCD-HCP pipeline includes two additional modifications to improve output image quality. ANTs 65 DenoiseImage models scanner noise as a Rician distribution and attempts to remove such noise from the T1 and T2 anatomical images. Additionally, ANTs N4BiasFieldCorrection attempts to smooth relative image histograms in different parts of the brain and improves bias field correction. 2) FreeSurfer 1 constructs cortical surfaces from the normalized anatomical data. This stage performs anatomical segmentation, white/grey and grey/CSF cortical surface construction, and surface registration to a standard surface template. Surfaces are refined using the T2 weighted anatomical data. Mid-thickness surfaces, which represent the average of white/grey and grey/CSF surfaces, are generated here. 3) PostFreesurfer converts prior outputs into an HCP-compatible format (i.e. CIFTIs) and transforms the volumes to a standard volume template space using ANTs nonlinear registration, and the surfaces to the standard surface

	space via spherical registration.
Normalization template	The "Vol" stage corrects for functional distortions via reverse-phase encoding spin-echo images. All resting state runs underwent intensity normalization to a whole brain mode value of 1000, within run correction for head movement, and functional data registration to the standard template (MNI). Atlas transformation was computed by registering the mean intensity image from each BOLD session to the high resolution T1 image, and then applying the anatomical registration to the BOLD image. This atlas transformation, mean field distortion correction, and resampling to 3-mm isotropic atlas space were combined into a single interpolation using FSL's 66 applywarp tool. The "Surf" stage projects the normalized functional data onto the template surfaces.
Noise and artifact removal	Additional BOLD preprocessing steps were executed to reduce spurious variance unlikely to reflect neuronal activity 46. First, a respiratory filter was used to improve FD estimates calculated in the volume ("vol") stage68. Second, temporal masks were created to flag motion-contaminated frames using the improved FD estimates 63. Frames with a filtered FD>0.3mm were flagged as motion-contaminated for nuisance regression only. After computing the temporal masks for high motion frame censoring, the data were processed with the following steps: (i) demeaning and detrending, (ii) interpolation across censored frames using least squares spectral estimation of the values at censored frames so that continuous data can be (iii) denoised via a GLM with whole brain, ventricular, and white matter signal regressors, as well as their derivatives. Denoised data were then passed through (iv) a band-pass filter (0.008 Hz<f<0.10 Hz) without re-introducing nuisance signals 69 or contaminating frames near high motion frames.
Volume censoring	Yes, ABCD data were censored at a filtered frame-wise displacement of < 0.08mm and HCP data were filtered using a non-filtered framewise displacement of <0.20mm.

Statistical modeling & inference

Model type and settings	Mass univariate and multivariate (support vector regression, canonical correlation analysis). Multiple parameterizations of each of these models were explored with the stated goal being to determine field-wide reproducibility in brain-phenotype association studies (see manuscript).
Effect(s) tested	As the primary aim of the paper was to determine the general reproducibility of brain-phenotype effects, multiple scales and combinations of effects were examined. Owing to the cross-sectional, nature of these studies, all effects are between-person associations.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☒ Both

Anatomical location(s)	Parcel-level and network-level analyses utilized the field-standard Gordon et al., 2016, Cerebral Cortex, and Seitzman et al., 2020, NeuroImage. Vertex-wise and voxel-wise data were extracted from Ciftis.
Statistic type for inference (See Eklund et al. 2016)	Multiple levels of neuroanatomical scale were used, including voxels, regions of interest, and networks.
Correction	As the primary aim of the paper was to determine the general reproducibility of brain-phenotype effects, multiple levels of significance values and correction were used, ranging from uncorrected to bonferroni (FWER) correction.

Models & analysis

n/a	Involved in the study
☐	☒ Functional and/or effective connectivity
☒	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Functional and/or effective connectivity	Pearson correlation
Multivariate modeling and predictive analysis	Two supervised regression models were used: a Ridge Regression model (α = 1.0), as proposed by Spisak et al. and a combined Principal Component Analysis (PCA) and Support Vector Regression (SVR) model, whereby half of the principal components (retaining 50% of the variance) generated from the PCA were passed as features into the SVR, as in the original work by Marek, Tervo-Clemmens et al. Both models were implemented using scikit-learn 5 in Python 3. For both HCP and ABCD datasets, both methods (ridge regression; PCA+SVR) and using three different neuroimaging feature sets (RSFC: full correlation, partial correlation; cortical thickness), the same analyses were conducted using code directly from Spisak et al. (https://gitlab.com/DosenbachGreene/bwas_response). For each behavioural phenotype and neuroimaging feature set combination, in each dataset, a complete cases sub-dataset was compiled, removing participants with missing behavioural phenotypes or neuroimaging data. For each of these complete cases (per Spisak et al.) neuroimaging feature set behavioural phenotype sub-datasets, 100 bootstraps were run for each model. Within each bootstrap, the sub-dataset was equally and randomly split into a discovery and replication set based on a given sample size. Here, sample size is defined as the size of a sole discovery/training set (identical in size to the replication set), such that given a sample size n, the total number of participants/samples of the combined discovery and replication sets is 2n. Following Spisak et al. (method and code), the discovery set was divided again into 10 cross-validation folds. However, unlike the nested cross-validation which was explored in our original manuscript and shown to not substantively change results (Marek, Tervo-Clemmens et al. 1: Supplemental Fig. S11, S12), this procedure

utilised by Spisak et al., and repeated here, did not use the additional cross-validation step for hyperparameter tuning. Rather an additional out-of-sample test was applied to the discovery dataset. TThe analyses and Figures (Fig. 1, 2) in this work use combinations of Spisak et al.'s methodological suggestions and those from our original work to replicate, expand, and clarify Spisak et al.'s Matters Arising commentary and to provide a more comprehensive perspective on out-of-sample multivariate BWAS effects. Rationale and additional details for specific analyses are provided in the relevant "Main Text" and "Figure Captions". In all cases, out-of-sample associations were evaluated as the correlation between the predicted phenotype score and the true score in the out-of-sample data. In-sample (training) associations were evaluated as the correlation between the true score and the predicted score from the model developed in the discovery set (that is, the data in the sample used to develop the model (Fig. 1).

Successful out-of-sample replication was defined as in Spisak et al.: 80% of bootstrapped iterations for a given behavioural phenotype-brain feature set ("BWAS") that were significant (via permutation test) in the first cross-validation test are significant in the second, split half test. We note this definition of replication by Spisak et al. thus does not consider all bootstrap iterations (n = 100) run when determining replication success/failure. That is, the denominator of a replication percentage is set by the number of bootstrap iterations that are significant in the first cross-validation test. Therefore, to ensure this measure of 80% replication represented a true percentage, replication here also required that more than one bootstrap iteration (out of the total 100) replicated (as defined above). Without this criteria, the impact of sampling variability and the performance of a single bootstrap iteration ensured that a small number of BWAS would appear to intermittently have replication successes followed by replication failure for the very smallest sample sizes. Reproducibility estimates following Spisak et al. guidelines were highly consistent with those from our original work.

Matters arising

Multivariate BWAS can be replicable with moderate sample sizes

https://doi.org/10.1038/s41586-023-05745-x

Received: 10 April 2022

Accepted: 19 January 2023

Published online: 8 March 2023

Open access

Check for updates

Tamas Spisak[1,2]✉, Ulrike Bingel[2] & Tor D. Wager[3]

ARISING FROM: S. Marek et al. Reproducible brain-wide association studies require thousands of individuals. *Nature* https://doi.org/10.1038/s41586-022-04492-9 (2022)

Brain-wide association studies (BWAS)—which correlate individual differences in phenotypic traits with measures of brain structure and function—have become a dominant method for linking mind and brain over the past 30 years. Univariate BWAS typically test tens to hundreds of thousands of brain voxels individually, whereas multivariate BWAS integrate signals across brain regions into a predictive model. Numerous problems have been raised with univariate BWAS, including a lack of power and reliability and an inability to account for pattern-level information embedded in distributed neural circuits[1–4]. Multivariate predictive models address many of these concerns, and offer substantial promise for delivering brain-based measures of behavioural and clinical states and traits[2,3].

In their recent paper[4], Marek et al. evaluated the effects of sample size on univariate and multivariate BWAS in three large-scale neuroimaging datasets and came to the general conclusion that "BWAS reproducibility requires samples with thousands of individuals". We applaud their comprehensive analysis, and we agree that (1) large samples are needed when conducting univariate BWAS and (2) multivariate BWAS reveal substantially larger effects and are therefore more highly powered.

Marek et al.[4] find that multivariate BWAS provide inflated in-sample associations that often cannot be replicated (that is, are underpowered) unless thousands of participants are included. This implies that effect-size estimates from the discovery sample are necessarily inflated. However, we distinguish between the effect-size estimation method (in-sample versus cross-validated) and the sample (discovery versus replication), and show that, with appropriate cross-validation, the in-sample inflation that Marek et al.[4] report in the discovery sample can be entirely eliminated. With additional analyses, we demonstrate that multivariate BWAS effects in high-quality datasets can be replicable with substantially smaller sample sizes in some cases. Specifically, applying a standard multivariate prediction algorithm to functional connectivity in the Human Connectome Project yielded replicable effects with sample sizes of 75–500 for 5 of 6 phenotypes tested (Fig. 1).

These analyses are limited to a selected number of phenotypes in a relatively high-quality dataset (measured in a young adult population with a single scanner) and should not be overgeneralized. However, they highlight that the key determinant of sample size requirements is the true effect size of the brain–phenotype relationship and that, with proper internal validation, appropriate effect-size estimates and sufficiently large effects for moderately sized studies are possible.

Marek et al.[4] evaluate in-sample effect-size inflation in multivariate BWAS by training various multivariate models in a 'discovery sample' and comparing the in-sample effect sizes (prediction–outcome correlation, r) estimated from the training sample to the performance in an independent replication sample. On the basis of a bootstrap analysis, with variously sized pairs of samples drawn randomly from the Adolescent Brain Cognitive Development study, the authors report a severe effect-size inflation of $\Delta r = -0.29$ (average difference between the in-sample effect sizes in the discovery sample and the out-of-sample effect sizes in the replication sample) and conclude that "[e]ven at the largest sample sizes ($n \approx 2{,}000$), multivariate in-sample associations remained inflated on average".

The issue with claims of inflation is that the in-sample effect size estimates of Marek et al.[4] were based on training multivariate models on the entire discovery sample, without cross-validation or other internal validation (as confirmed by inspection of the code and discussion with the authors). Such in-sample correlations are not valid effect-size estimates, as they produce a well-known overfitting bias that increases with model complexity[5]. Standard practice in machine learning is to evaluate model accuracy (and other performance metrics) on data independent of those used for training. In line with current recommendations for multivariate brain–behaviour analyses[6,7], this is typically performed using internal cross-validation (for example, k-fold) to estimate unbiased effect sizes in a discovery sample, and (less commonly) further validating significant cross-validated effects in held-out or subsequently acquired replication samples[2,5].

Using cross-validation to estimate discovery-sample effects impacts the pool of studies selected for replication attempts, the degree of effect-size attenuation in replication samples, and the sample size needed for effective replication and mitigation of publication bias. To demonstrate this and provide quantitative estimates of sample size requirements in multivariate BWAS, we analysed functional connectivity data from the Human Connectome Project[8] (one of the datasets in Marek et al.[4]) using cross-validation to estimate discovery-sample effect sizes. As shown in Fig. 1a–d, cross-validated discovery effect-size estimates are unbiased (that is, not inflated on average), irrespective of the sample size and the magnitude of the effect. As expected, even with cross-validation, smaller sample sizes resulted in lower power (Fig. 1e) and increased variability in effect-size estimates across samples (Fig. 1c). Such variability is undesirable because it reduces the probability of independent replication (Fig. 1f). Moreover, selection biases—most notably, publication bias—can capitalize on such variability to inflate effect sizes in the literature (Fig. 1g). Although these effects of using small sample sizes are undesirable, they do not invalidate

[1]Institute of Diagnostic and Interventional Radiology and Neuroradiology, University Medicine Essen, Essen, Germany. [2]Center for Translational Neuro- and Behavioral Sciences, Department of Neurology, University Medicine Essen, Essen, Germany. [3]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA. ✉e-mail: tamas.spisak@uk-essen.de

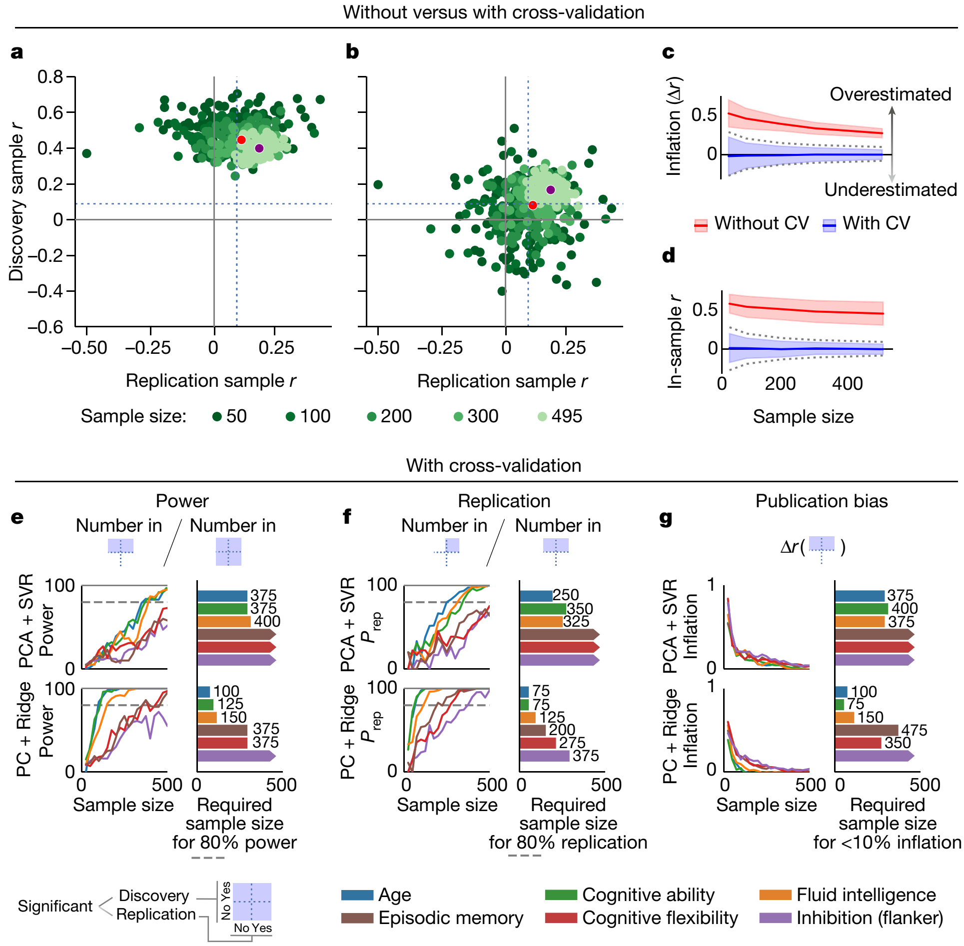


Fig. 1 | Examples of multivariate BWAS providing unbiased effect sizes and high replicability with low to moderate sample sizes. a, Discovery sample effects in multivariate BWAS are inflated only if estimates are obtained without cross-validation (CV). **b**, Cross-validation fully eliminates in-sample effect-size inflation and, as a consequence, provides higher replicability. Data are from the Human Connectome Project (HCP1200, PTN release, $n = 1{,}003$). Each point in **a** and **b** corresponds to one bootstrap subsample, as in figure 4b of Marek et al.[4]. The dotted lines denote the threshold for $P = 0.05$ with $n = 495$. Mean multivariate brain–behavioural phenotype associations across 100 bootstrap samples at $n = 200$ and for the full sample are denoted by red and purple dots. **c**, The inflation of in-sample effect size obtained without cross-validation (red) is reduced, but does not disappear, at higher sample sizes. Conversely, cross-validated estimates (blue) are slightly pessimistic with low sample sizes and become quickly unbiased as sample size is increased. **d**, Without cross-validation, in-sample effect-size estimates are non-zero ($r \approx 0.5$, red), even when predicting permuted outcome data. Cross-validation eliminates systematic bias across all sample sizes (blue). The dashed lines in **c** and **d** denote 95% parametric confidence intervals, and the shaded areas denote bootstrap- and permutation-based confidence intervals. **e**,**f**, Cross-validated analysis reveals that sufficient in-sample power (**e**) and out-of-sample replication probability (P_{rep}) (**f**) can be achieved for a variety of phenotypes at low or moderate sample sizes. 80% power and P_{rep} are achievable in <500 participants for 3 out of 6 phenotypes (coloured bars) using the prediction algorithm of Marek et al.[4] (**e** and **f** (top), the sample size required for 80% power or P_{rep} is shown). The remaining three phenotypes require sample sizes of >500 (bars with arrows). Power and P_{rep} can be substantially improved with a ridge regression-based model recommended in some comparison studies[10,11] (**e** and **f** (bottom), with 80% power and P_{rep} with sample sizes as low as $n = 100$ and $n = 75$, respectively, when predicting cognitive ability, and sample sizes between 75 and 375 for other investigated variables (fluid intelligence, episodic memory and cognitive flexibility), except inhibition assessed with the flanker task, which replicated with $n = 375$ but did not reach 80% power with $n = 500$. **g**, We estimated interactions between sample size and publication bias by computing effect size inflation ($r_{discovery} - r_{replication}$) only for those bootstrap cases in which prediction performance was significant ($P > 0.05$) in the replication sample. Our analysis shows that the effect-size inflation due to publication bias is modest (<10%) with fewer than 500 participants for half of the phenotypes using the model from Marek et al.[4] and all phenotypes but the flanker using the ridge model. The blue squares show conditional relationships assessed to derive metrics in **e**, **f** and **g** with reference to **b**. The top and bottom squares indicate positive and negative results in the discovery sample, respectively. The left and right squares indicate negative and positive results in the replication sample. The blue squares indicate how these conditions were applied to derive the metrics.

the use of multivariate BWAS in small samples, and publication biases can be mitigated by practices that, like internal cross-validation, are quickly becoming standards in the field[2,5]. These include preregistration, registered reports, reporting confidence intervals and the use of hold-out samples tested only once on a single, optimized model to avoid overfitting.

Given these considerations, we wondered how many participants are generally required for multivariate BWAS. The answer to this question depends on the reliability of both phenotypic and brain measures, the size of the effects linking them, the algorithm and model-selection steps used and the use cases for the resulting brain measures. For example, multivariate models trained on as few as 20 participants[9] can have high reliability (ICC = 0.84)[10], broad external validity and large effect sizes (Hedges $g = 2.3$)[11] in independent samples (for example, more than 600 participants from 20 independent studies in ref. [11]) when predicting behavioural states within-person rather than traits. In this case, the benefit of large samples is primarily in accurately estimating local brain weights (model parameters)[12] rather than increasing out-of-sample

accuracy. Here we performed functional connectivity-based multivariate BWAS with cognitive ability (the phenotype shown in figure 4 of Marek et al.[4]) and five other cognition-related example phenotypes selected at random and demonstrate that, even when predicting trait-level phenotypes, as Marek et al.[4] did, sample sizes of 75–500 are sufficient in five out of six of cases that we tested (or three out of six cases using the prediction algorithm of Marek et al.[4]) to achieve high statistical power and replicability (for example, 80%) and to mitigate effect size inflation due to publication bias.

The basis for these estimates is shown in Fig. 1e–g. Using cross-validated discovery sample effect-size estimates, the multivariate BWAS model of Marek et al.[4]—principal-component-based reduction of bivariate connectivity followed by support vector regression (PCA + SVR)—showed 80% in-sample power and 80% out-of-sample replication probability (P_{rep}) at $n < 500$ for three out of six phenotypes that we examined (age, cognitive ability and fluid intelligence). However, this model has been shown to be disadvantageous in some comparison studies[12,13]. We therefore performed the same power and sample-size calculations for a multivariate BWAS using another approach—ridge regression on partial correlation matrices with a default shrinkage parameter of 1 (PC + ridge; Supplementary Methods). Although this approach is still probably suboptimal[12,13] (we avoided testing other models to avoid overfitting), it substantially improved the power (Fig. 1e (bottom)), independent replication probability ([P_{rep}]; Fig. 1f (bottom)) and resistance to inflation due to publication bias (Fig. 1g (bottom)). Eighty per cent power and P_{rep} were achieved at sample sizes from 75 to 150 for age (included as a reference variable), cognitive ability and fluid intelligence, and sample sizes < 400 for all phenotypes except for inhibition measured by the flanker task (a measure that is known to have low reliability[14]).

Our results highlight, that the key determinant of sample size requirements is the true effect size of the brain–phenotype relationship, which subsumes the amount, quality, homogeneity and reliability of both brain and phenotypic measures, and the degree to which a particular brain measure is relevant to a particular phenotype. Effect sizes will probably vary widely across studies; for example, cortical thickness can also reliably predict 4 out of the 6 investigated phenotypes with $n < 500$, although with smaller effect sizes on average (functional connectivity, mean $r = 0.2$; cortical thickness, mean $r = 0.1$; Supplementary Fig. 2). Although our results were derived from a relatively high-quality dataset and used an algorithm expected to yield larger effect sizes than that of Marek et al.[4], they are in agreement with analytical calculations showing that BWAS that explain more than 1% of the phenotype's variance can be replicable with sample sizes below 1,000 (Supplementary Methods). For example, a model that explains $r^2 = 0.01$ (1% of variance) achieves 80% power in a prospective replication with $n = 801$, and $r^2 = 0.02$ achieves 80% power with $n = 399$ (ref. [15]).

These quantitative differences in required sample size could translate into large, qualitative differences in the types of neuroimaging studies considered viable in future efforts. There is a necessary trade-off between the innovativeness of a task, measure or method, and the extent to which it has been validated. Existing large-scale neuroimaging studies ($n > 1{,}000$) have selected well-validated tasks and imaging measures over new, exploratory ones, and few have attempted to characterize rare populations. Requiring sample sizes that are larger than necessary for the discovery of new effects could stifle innovation.

We agree with Marek et al.[4] that small-sample studies are important for understanding the brain bases of tasks and mental states[9–11], and for prototyping new tasks and measures. Furthermore, several current trends may further increase the viability of small-sample multivariate BWAS, including (1) new phenotypes, (2) feature-learning methods and algorithms with larger effect sizes[13], (3) models that target within-person variation in symptoms and behaviour to improve between-person predictions[2] and (4) hybrid strategies for improving prediction like meta-matching[16]. All of these have the potential to improve reliability and effect sizes, but whether they do remains to be seen.

Finally, as both Marek et al.[4] and our analyses show, very small effects will suffer from limited power, replicability and predictive utility even with sample sizes in the thousands (Fig. 1). We argue that the field should focus on discovering phenotypes and brain measures with large effect sizes. Efficient discovery entails casting a wide net in smaller studies using rigorous, unbiased methods and scaling up promising findings to larger samples[2]. There are substantial challenges ahead, including establishing broad generalizability across contexts, equity across subpopulations, and models with high neuroscientific validity and interpretability[17,18]. Addressing these challenges will require innovative new methods and measures.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-05745-x.

Reporting summary

Further information on experimental design is available in the Nature Portfolio Reporting Summary linked to this Article.

Data availability

Analysis is based on preprocessed data provided by the Human Connectome Project, WU-Minn Consortium (principal investigators: D. Van Essen and K. Ugurbil; 1U54MH091657) funded by the 16 NIH institutes and centres that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. All data used in the present study are available for download from the Human Connectome Project (www.humanconnectome.org). Users must agree to data use terms for the HCP before being allowed access to the data and ConnectomeDB; details are provided online (https://www.humanconnectome.org/study/hcp-young-adult/data-use-terms).

Code availability

All analysis code used in the current study is available at GitHub (release v.1.0; https://github.com/spisakt/BWAS_comment).

1. Button, K. S. et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
2. Woo, C.-W. W., Chang, L. J., Lindquist, M. A. & Wager, T. D. Building better biomarkers: Brain models in translational neuroimaging. *Nat. Neurosci.* **20**, 365–377 (2017).
3. Bzdok, D., Varoquaux, G. & Steyerberg, E. W. Prediction, not association, paves the road to precision medicine. *JAMA Psychiatry* **78**, 127–128 (2021).
4. Marek, S. et al. Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654–660 (2022).
5. Poldrack, R. A., Huckins, G. & Varoquaux, G. Establishment of best practices for evidence for prediction: a review. *JAMA Psychiatry* **77**, 534–540 (2020).
6. Genon, S., Eickhoff, S. B. & Kharabian, S. Linking interindividual variability in brain structure to behaviour. *Nat. Rev. Neurosci.* **23**, 307–318 (2022).
7. Rosenberg, M. D. & Finn, E. S. How to establish robust brain–behavior relationships without thousands of individuals. *Nat. Neurosci.* **25**, 835–837 (2022).
8. Van Essen, D. C. et al. The WU-Minn human connectome project: an overview. *Neuroimage*, **80**, 62–79 (2013).
9. Wager, T. D. et al. An fMRI-based neurologic signature of physical pain. *N. Engl. J. Med.* **368**, 1388–1397 (2013).
10. Han, X. et al. Effect sizes and test-retest reliability of the fMRI-based neurologic pain signature. *Neuroimage* **247**, 118844 (2022).
11. Zunhammer, M., Bingel, U. & Wager, T. D. Placebo effects on the neurologic pain signature. *JAMA Neurol.* **75**, 1321–1330 (2018).
12. Tian, Y. & Zalesky, A. Machine learning prediction of cognition from functional connectivity: are feature weights reliable? *Neuroimage* **245**, 118648 (2021).
13. Pervaiz, U., Vidaurre, D., Woolrich, M. W. & Smith, S. M. Optimising network modelling methods for fMRI. *Neuroimage* **211**, 116604 (2020).
14. Hedge, C., Powell, G. & Sumner, P. The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods* **50**, 1166–1186 (2018).

15. Killeen, P. R. Predict, control, and replicate to understand: how statistics can foster the fundamental goals of science. *Perspect. Behav. Sci.* **42**, 109–132 (2018).
16. He, T. et al. Meta-matching as a simple framework to translate phenotypic predictive models from big to small data. *Nat. Neurosci.* **25**, 795–804 (2022).
17. Wu, J. et al. Cross-cohort replicability and generalizability of connectivity-based psychometric prediction patterns. *Neuroimage* **262**, 119569 (2022).
18. Li, J. et al. Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Sci. Adv.* **8**, eabj1812 (2022).

Acknowledgements We thank S. Marek et al. for sharing the analysis code and for the discussions in relation to our commentary. The work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; projects 'TRR289 - Treatment Expectation', ID 422744262 and 'SFB1280 - Extinction Learning', ID 316803389), R01 MH076136 and R01 EB026549.

Author contributions Conception and data analysis: T.S. Manuscript writing and revision: T.S., U.B. and T.D.W.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-05745-x.

Correspondence and requests for materials should be addressed to Tamas Spisak.

Reprints and permissions information is available at http://www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

nature portfolio

Corresponding author(s): Tamas Spisak

Last updated by author(s): Dec 20, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Analysis is based on preprocessed data provided by the Human Connectome Project, WU-Minn Consortium (principal investigators: D. Van Essen and K. Ugurbil; 1U54MH091657) funded by the 16 NIH institutes and centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. All data used in the present study are available for download from the Human Connectome Project (www.humanconnectome.org). Users must agree to data use terms for the HCP before being allowed access to the data and ConnectomeDB; details are provided at https://www.humanconnectome.org/study/hcp-young-adult/data-use-terms. Preprocessed data was created software as described in: Glasser MF, Sotiropoulos SN, Wilson JA, Coalson TS, Fischl B, Andersson JL, Xu J, Jbabdi S, Webster M, Polimeni JR, Van Essen DC. The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage. 2013 Oct 15;80:105-24.
Data analysis	https://github.com/spisakt/BWAS_comment v1.0 Dependencies: python 3.10.8 numpy 1.23.5 pandas 1.5.2 scikit-learn 1.2.0 joblib 1.2.0 mlxtend 0.21.0 seaborn 0.12.1 matplotlib 3.6.2

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Analysis is based on preprocessed data provided by the Human Connectome Project, WU-Minn Consortium (principal investigators: D. Van Essen and K. Ugurbil; 1U54MH091657) funded by the 16 NIH institutes and centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. All data used in the present study are available for download from the Human Connectome Project (www.humanconnectome.org). Users must agree to data use terms for the HCP before being allowed access to the data and ConnectomeDB; details are provided at https://www.humanconnectome.org/study/hcp-young-adult/data-use-terms.
All derivative data, (including raw material for figures) is freely available at https://github.com/spisakt/BWAS_comment

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	The Human Connectome Project involved 656 females and 550 males. Findings apply to both males and females. Sex and gender differences are out of the scope of the Matters Arising and were not considered in the analysis. For more information, refer to van Essen et al., 2013.
Population characteristics	Refer to van Essen et al., 2013.
Recruitment	As described in van Essen et al., 2013.
Ethics oversight	As described in van Essen et al., 2013.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Quantitative analyses of the replicability of functional connectivity-based brain-wise association studies.
Research sample	We re-analyzed one of the studies involved in the original publication by Marek, Tervo-Clemmens et al., including open access-data from the Human Connectome Project (van Essen et al., 2013), based on a non-representative sample of young adults (656 females, 550 males, mean+-sd age: 28.9+-3.57).
Sampling strategy	We re-analyzed one of the studies involved in the original publication by Marek, Tervo-Clemmens et al., including open access-data from the Human Connectome Project. For more information, please refer to van Essen et al., 2013 ands Marek, Tervo-Clemmens et al, 2022.
Data collection	See van Essen et al., 2013 ands Marek, Tervo-Clemmens et al, 2022 for details.
Timing	See van Essen et al., 2013 ands Marek, Tervo-Clemmens et al, 2022 for details.
Data exclusions	We excluded participants with no MRI images available, and for each analysis, participants with missing data about the target phenotype.
Non-participation	203 out of the 1206 HCP participants didn't have MRI data. Exclusion die to missing phenotype data varied across analyses (less than 20 in all cases)
Randomization	N/A (observational study)

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Magnetic resonance imaging

Experimental design

Design type	resting state fMRI, anatomical MRI
Design specifications	4, 15min, eyes open (see van Essen et al., 2013)
Behavioral performance measures	As described in van Essen et al., 2013.

Acquisition

Imaging type(s)	functional, anatomical
Field strength	3
Sequence & imaging parameters	As described in van Essen et al., 2013.
Area of acquisition	brain

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software	As described in: Glasser MF, Sotiropoulos SN, Wilson JA, Coalson TS, Fischl B, Andersson JL, Xu J, Jbabdi S, Webster M, Polimeni JR, Van Essen DC. The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage. 2013 Oct 15;80:105-24.
Normalization	As described in: van Essen et al., 2013.
Normalization template	As described in: van Essen et al., 2013.
Noise and artifact removal	As described in: van Essen et al., 2013.
Volume censoring	As described in: van Essen et al., 2013.

Statistical modeling & inference

Model type and settings	Multivariate predictive modelling, PCA+SVR and Ridge, cross-validation (see Supplementary Methods for details).
Effect(s) tested	Predictive performance (Pearson's r)

Specify type of analysis: ☒ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016)	N/A (predictive modelling)
Correction	N/A (predictive modelling)

Models & analysis

n/a	Involved in the study
☐	☒ Functional and/or effective connectivity
☒	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Functional and/or effective connectivity	partial correlation
Multivariate modeling and predictive analysis	Targets: age, cognitive ability, episodic memory, fluid intelligence, cognitive flexibility, inhibition. Model: PCA +SVR (reproduction of Marek et al.'s model), Ridge regression (with the default hyperparameter value 1). The PCA+SVR model involved dimensionality reduction. Features partial correlation values across 100 ICA-based regions, training metric: mean squared error, evaluation metric, mean squared error and Pearson correlation.

Reply to: Multivariate BWAS can be replicable with moderate sample sizes

https://doi.org/10.1038/s41586-023-05746-w
Published online: 8 March 2023
Open access
Check for updates

Brenden Tervo-Clemmens[1,22]✉, Scott Marek[2,3,22]✉, Roselyne J. Chauvin[4], Andrew N. Van[4,5], Benjamin P. Kay[4], Timothy O. Laumann[3], Wesley K. Thompson[6], Thomas E. Nichols[7], B. T. Thomas Yeo[8,9,10,11,12,13], Deanna M. Barch[3,14], Beatriz Luna[15,16], Damien A. Fair[17,18,19,23]✉ & Nico U. F. Dosenbach[2,4,5,14,20,21,23]✉

REPLYING TO: T. Spisak et al. *Nature* https://doi.org/10.1038/s41586-023-05745-x (2023)

In our previous study[1], we documented the effect of sample size on the reproducibility of brain-wide association studies (BWAS) that aim to cross-sectionally relate individual differences in human brain structure (cortical thickness) or function (resting-state functional connectivity (RSFC)) to cognitive or mental health phenotypes. Applying univariate and multivariate methods (for example, support vector regression (SVR)) to three large-scale neuroimaging datasets (total $n \approx 50{,}000$), we found that overall BWAS reproducibility was low for $n < 1{,}000$, due to smaller than expected effect sizes. When samples and true effects are small, sampling variability, and/or overfitting can generate 'statistically significant' associations that are likely to be reported due to publication bias, but are not reproducible[2–5], and we therefore suggested that BWAS should build on recent precedents[6,7] and continue to aim for samples in the thousands. In the accompanying Comment, Spisak et al.[8] agree that larger BWAS are better[5,9], but argue that "multivariate BWAS effects in high-quality datasets can be replicable with substantially smaller sample sizes in some cases" (n = 75–500); this suggestion is made on the basis of analyses of a selected subset of multivariate cognition/RSFC associations with larger effect sizes, using their preferred method (ridge regression with partial correlations) in a demographically more homogeneous, single-site/scanner sample (Human Connectome Project (HCP), n = 1,200, aged 22–35 years).

There is no disagreement that a minority of BWAS effects can replicate in smaller samples, as shown with our original methods[1]). Using the exact methodology (including cross-validation) and code of Spisak et al.[8] to repeat 64 multivariate BWAS in the 21-site, larger and more diverse Adolescent Brain Cognitive Development Study (ABCD, n = 11,874, aged 9–11 years), we found that 31% replicated at n = 1,000, dropping to 14% at n = 500 and none at n = 75. Contrary to the claims of Spisak et al.[8], replication failure was the outcome in most cases when applied to this larger, more diverse dataset. Basing general BWAS sample size recommendations on the largest effects has at least two fundamental flaws: (1) failing to detect other true effects (for example, reducing the sample size from n = 1,000 to n = 500 leads to a 55% false-negative rate), therefore restricting BWAS scope, and (2) inflation of reported effects[3,10–12]. Thus, regardless of the method, associations based on small samples can remain distorted and lack generalizability until confirmed in large, diverse, independent samples.

We always test for BWAS replication with null models (using permutation tests) of out-of-sample estimates to ensure that our reported reproducibility is unaffected by in-sample overfitting. Nonetheless, Spisak et al.[8] argue against plotting inflated in-sample estimates[1,10] on the y axis, and out-of-sample values on the x axis, as we did (Fig. 1a). Instead, they propose plotting cross-validated associations from an initial, discovery sample (Fig. 1b (y axis)) against split-half out-of-sample associations (x axis). However, cross-validation—just like split-half validation—estimates out-of-sample, and not in-sample, effect sizes[13]. The in-sample associations[1,10] for the method of Spisak et al.[8] (Fig. 1b), that is, from data in the sample used to develop the model, show the same degree of overfitting (Fig. 1a versus Fig. 1b). The plot of Spisak et al.[8] (Fig. 1c) simply adds an additional out-of-sample test (cross-validation before split half), and therefore demonstrates the close correspondence between two different methods for out-of-sample effect estimation[14]. Analogously, we can replace the cross-validation step in the code of Spisak et al.[8] with split-half validation (our original out-of-sample test), obtaining split-half effects in the first half of the sample, and then comparing them to the split-half estimates from the full sample (Fig. 1d). The strong correspondences between cross-validation followed by split-half (Spisak et al. method[8]; Fig. 1c) and repeated split-half validation (Fig. 1d) are guaranteed by plotting out-of-sample estimates (from the same dataset) against one another. Here, plotting cross-validated discovery sample estimates on the y axis (Fig. 1c,d) provides no additional information beyond the x axis out-of-sample values. The critically important out-of-sample predictions, required for reporting multivariate results[1], generated using the method of Spisak et al.[8] and our method are nearly identical (Fig. 1e).

As Spisak et al.[8] highlight, cross-validation of some type is considered to be standard practice[10], and yet the distribution of out-of-sample

[1]Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [2]Department of Radiology, Washington University School of Medicine, St Louis, MO, USA. [3]Department of Psychiatry, Washington University School of Medicine, St Louis, MO, USA. [4]Department of Neurology, Washington University School of Medicine, St Louis, MO, USA. [5]Department of Biomedical Engineering, Washington University in St Louis, St Louis, MO, USA. [6]Division of Biostatistics, University of California San Diego, La Jolla, CA, USA. [7]Oxford Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of Oxford, Oxford, UK. [8]Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore. [9]Centre for Sleep and Cognition, National University of Singapore, Singapore, Singapore. [10]Centre for Translational MR Research, National University of Singapore, Singapore, Singapore. [11]N.1 Institute for Health, Institute for Digital Medicine, National University of Singapore, Singapore, Singapore. [12]Integrative Sciences and Engineering Programme, National University of Singapore, Singapore, Singapore. [13]Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA, USA. [14]Department of Psychological and Brain Sciences, Washington University in St Louis, St Louis, MO, USA. [15]Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA. [16]Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA. [17]Masonic Institute for the Developing Brain, University of Minnesota Medical School, Minneapolis, MN, USA. [18]Department of Pediatrics, University of Minnesota Medical School, Minneapolis, MN, USA. [19]Institute of Child Development, University of Minnesota Medical School, Minneapolis, MN, USA. [20]Program in Occupational Therapy, Washington University School of Medicine, St Louis, MO, USA. [21]Department of Pediatrics, Washington University School of Medicine, St Louis, MO, USA. [22]These authors contributed equally: Brenden Tervo-Clemmens, Scott Marek. [23]These authors jointly supervised this work: Damien A. Fair, Nico U. F. Dosenbach. ✉e-mail: btervo-clemmens@mgh.harvard.edu; smarek@wustl.edu; faird@umn.edu; ndosenbach@wustl.edu

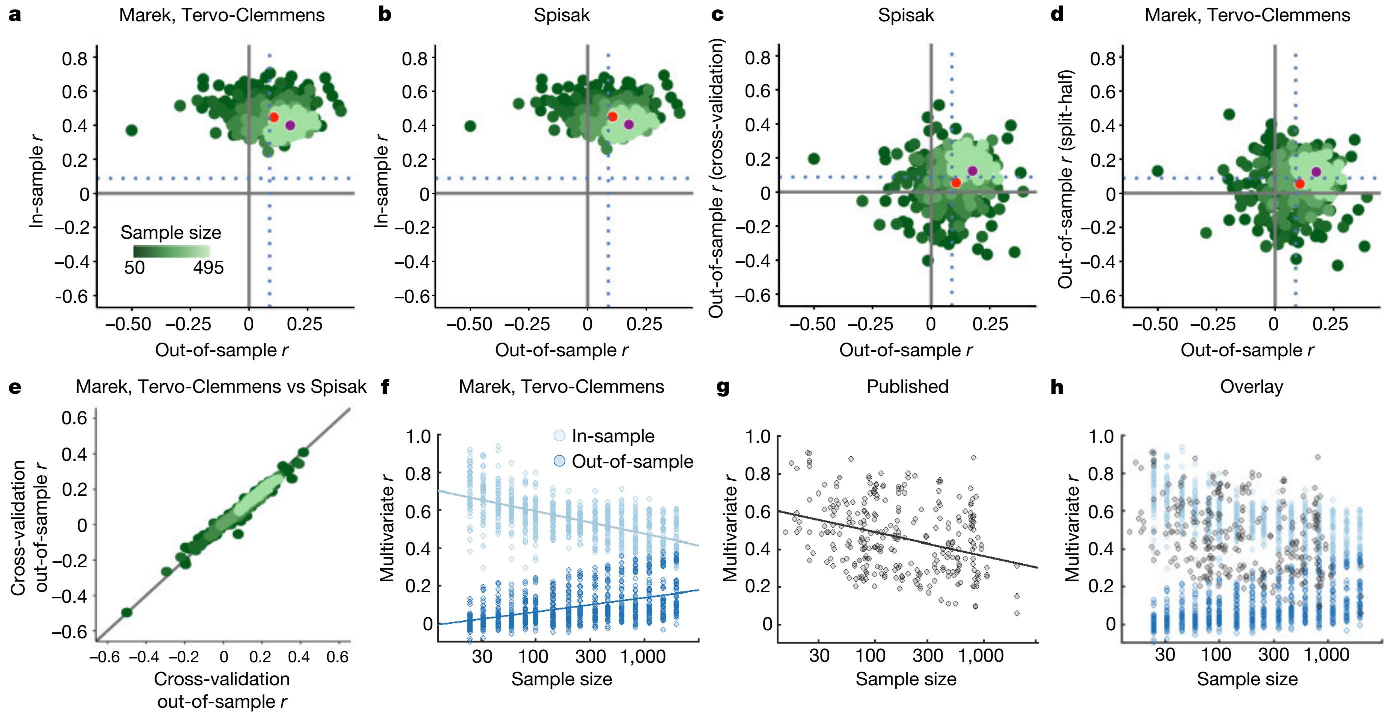


Fig. 1 | In-sample versus out-of-sample effect estimates in multivariate BWAS. a–**e**, Methods comparison between our previous study[1] (split-half) and Spisak et al.[8] (cross-validation followed by split-half). 'Marek, Tervo-Clemmens' and 'Spisak' refer to the methodolgies described in ref. [1] and ref. [8], respectively. For **a**–**e**, HCP 1200 Release (full correlation) data were used to predict age-adjusted total cognitive ability. Analysis code and visualizations (*x*,*y* scaling; colours) are the same as in Spisak et al.[8]. The *x* axes in **a**–**e** always display the split-half out-of-sample effect estimates from the second (replication) half of the data (correlation between true scores and predicted scores; as in Spisak et al.[8] and in our previous study[1]; Supplementary Methods). **a**, In-sample (training correlation; *y* axis) as a function of out-of-sample associations (plot convention in our previous study[1]). **b**, Matched comparison of the true in-sample association (training correlations, mean across folds; *y* axis) in the method proposed by Spisak et al.[8]. **c**, The proposed correction by Spisak et al.[8] that inserts an additional cross-validation step to evaluate the first half of data, which by definition makes this an out-of-sample association (*y* axis). **d**, Replacing the cross-validation step from Spisak et al.[8] with a split-half validation provides a different (compared with **c**) out-of-sample association of the first half of the total data (that is, each of the first stage split halves is one-quarter of the total data; *y* axis). The appropriate and direct comparison of in-sample associations between Spisak et al.[8] and our previous study[1] is comparing **b** to **a**, rather than **c** to **a**. The Spisak et al. method[8] (cross-validation followed by split-half validation) does not reduce in-sample overfitting (**b**) but, instead, adds an additional out-of-sample evaluation (**c**), which is nearly identical to split-half validation twice in a row (**d**), and makes it clear why the out-of-sample performance of these two methods is likewise nearly identical. **e**, Correspondence between out-of-sample associations (to the left-out half) from the additional cross-validation step proposed by Spisak et al.[8] (mean across folds; *y* axis) and the original split-half validation from our previous study[1] (*x* axis). The identity line is shown in black. **f**, In-sample (*r*; light blue) and out-of-sample (*r*; dark blue) associations as a function of sample size. Data are from figure 4a–d of ref. [1]. **g**, Published literature review of multivariate *r* (*y* axis) as a function of sample size (data from ref. [15]) displayed with permission. For **f** and **g**, best fit lines are displayed in $\log_{10}$ space. **h**, Overlap of **f** and **g**.

associations (Fig. 1f (dark blue)) does not match published multivariate BWAS results (Fig. 1g), which have largely ranged from $r = 0.25$ to 0.9, decreasing with increasing sample size[10,15,16]. Instead, published effects more closely follow the distribution of in-sample associations (Fig. 1h). This observation suggests that, in addition to small samples, structural problems in academic research (for example, non-representative samples, publication bias, misuse of cross-validation and unintended overfitting) have contributed to the publication of inflated effects[12,17,18]. A recent biomarker challenge[5] showed that cross-validation results continued to improve with the amount of time researchers spent with the data, and the models with the best cross-validation results performed worse on never-seen held-back data. Thus, cross-validation alone has proven to be insufficient and must be combined with the increased generalizability of large, diverse datasets and independent out-of-sample evaluation in new, never before seen data[5,10].

The use of additional cross-validation in the discovery sample by Spisak et al.[8] does not affect out-of-sample prediction accuracies (Fig. 1e). However, by using partial correlations and ridge regression on HCP data, they were able to generate higher out-of-sample prediction accuracies than our original results in ABCD (Fig. 2a). The five variables they selected are strongly correlated[19] cognitive measures from the NIH Toolbox (mean $r = 0.37$; compare with the correlation strength for height versus weight $r = 0.44$)[20] and age (not a complex behavioural phenotype), unrepresentative of BWAS as a whole (Fig. 2b (colour versus grey lines)). As the HCP is the relatively smallest and most homogeneous dataset, we applied the exact method and code of Spisak et al.[8] to the ABCD data (Fig. 2c and Supplementary Table 2). At $n = 1{,}000$ (training; $n = 2{,}000$ total), only 31% of BWAS (44% RSFC, 19% cortical thickness) were replicable (Fig. 2d; defined as in Spisak et al.[8]; Supplementary Information). Expanding BWAS scope beyond broad cognitive abilities towards complex mental health outcomes therefore requires $n > 1{,}000$ (Fig. 2b–d). The absolute largest BWAS (cognitive ability: RSFC, green) reached replicability only using $n = 400$ ($n = 200$ train; $n = 200$ test) with an approximate 40% decrease in out-of-sample prediction accuracies from HCP to ABCD (Fig. 2e (lighter green, left versus right)). The methods of Spisak et al.[8] and our previous study[1] returned equivalent out-of-sample reproducibility for this BWAS (cognitive ability: RSFC) in the larger, more diverse ABCD data (Fig. 2e (right, dark versus light green)). Thus, the smaller sample sizes (Fig. 2b,c) that are required for out-of-sample reproducibility (Fig. 2e) reported by Spisak et al.[8] in the HCP data did not generalize to the larger ABCD dataset. See also our previous study[1] for a broader discussion of convergent evidence across HCP and ABCD datasets.

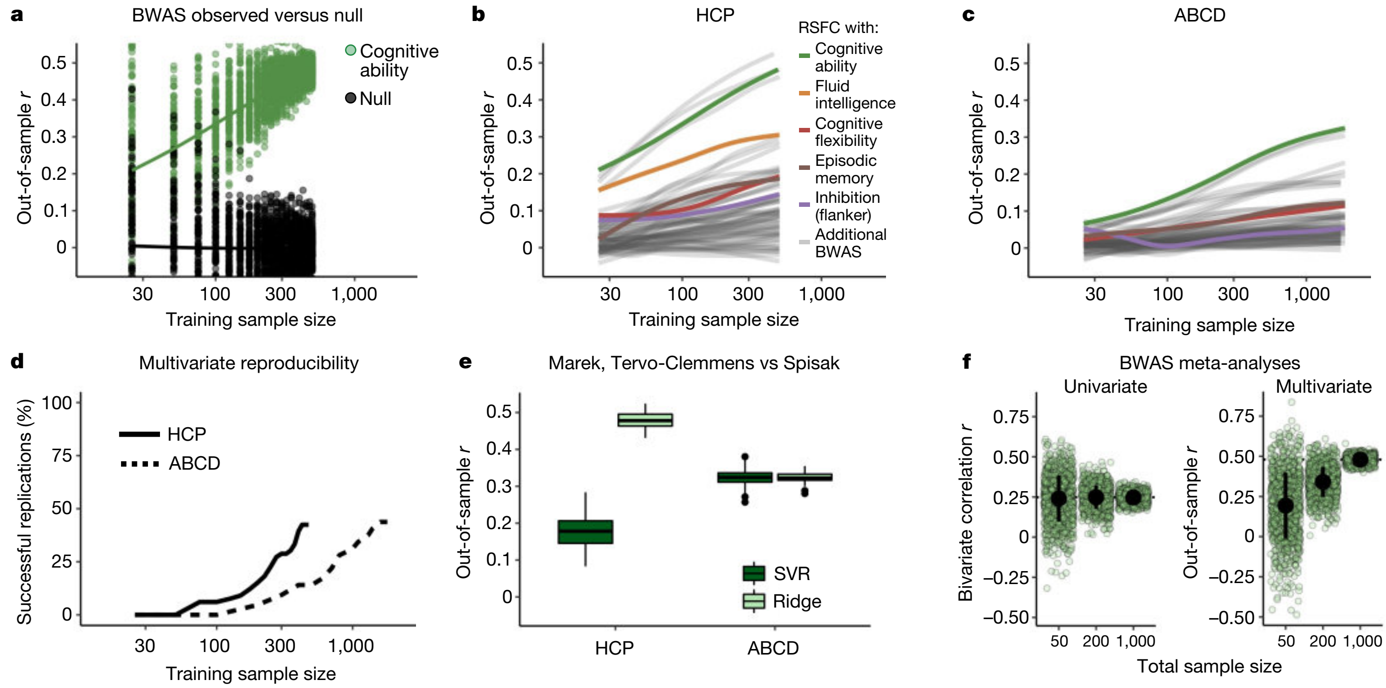


Fig. 2 | BWAS reproducibility, scope and prediction accuracy using the method of Spisak et al. a, Example bootstrapped BWAS of total cognitive ability (green) and null distribution (black) (*y* axis), as a function of sample size (*x* axis) from the suggested method of Spisak et al.[8] (RSFC by partial correlation; prediction by ridge regression) in the HCP dataset ($n = 1{,}200$, 1 site, 1 scanner, 60 min RSFC/participant, 76% white). Sample sizes were $\log_{10}$-transformed for visualization. **b**, Out-of-sample correlation (between true scores and predicted scores) from ridge regression (*y* axis; code from Spisak et al.[8]) as a function of training sample size (*x* axis, $\log_{10}$ scaling) for 33 cognitive and mental health phenotypes (Supplementary Information) in the HCP dataset. Each line displays a smoothed fit estimate (through penalized splines in general additive models) for a brain (RSFC (partial correlations, as proposed by Spisak et al.[8]), cortical thickness) phenotype pair (66 total) that has 100 bootstrapped iterations from sample sizes of 25 to 500 (inclusive) in increments of 25 (20 total bins). Sample sizes were $\log_{10}$-transformed (for visualization) before general additive model fitting. **c**, The same as in **b**, but in the ABCD dataset ($n = 11{,}874$, 21 sites, 3 scanner manufacturers, 20 min RSFC/participant, 56% white) using 32 cognitive and mental health phenotypes at sample sizes of 25, 50, 75 and from 100 to 1,900 (inclusive) in increments of 100 (22 total bins). **d**, The percentage of brain–phenotype pairs (BWAS) from **b** and **c** with significant replication on the basis of the method of Spisak et al.[8] (Supplementary Information). **e**, Comparison of our original method in our previous study[1] and the method proposed by Spisak et al.[8] at the full split-half sample size of HCP (left) and ABCD (right). Out-of-sample correlations (RSFC with total cognitive ability, *y* axis) for the method used in our previous study[1] (dark green; RSFC by correlation, PCA, SVR) and by Spisak et al.[8] (light green; RSFC by partial correlation, ridge regression). Repeating the method proposed by Spisak et al.[8] in ABCD (right) and comparing this to the method used in our previous study[1] results in a very similar out-of-sample *r*. **f**, Simulated individual studies (light green circles; $n = 1{,}000$ per sample size) and meta-analytic estimates (black dot, ±1 s.d.) using the method of Spisak et al.[8] (partial correlations in the HCP dataset) for the largest univariate association (left; *y* axis, bivariate correlation) and multivariate association (right; *y* axis, out-of-sample correlation) for total cognitive ability versus RSFC, as a function of total sample size (*x* axis; bivariate correlation for sample sizes of 50, 200 and 1,000, and multivariate sum of train and test samples, each 25, 100 and 500). For univariate approaches, studies of any sample size, when appropriately aggregated to a large total sample size, can correctly estimate the true effect size. However, for multivariate approaches, even when aggregating across 1,000 independent studies, studies with a small sample size produce prediction accuracies that are downwardly biased relative to large sample studies, highlighting the need for large samples in multivariate analyses.

Notably, the objections of Spisak et al.[8] raise additional reasons to stop the use of smaller samples in BWAS that were not highlighted in our original article. Multivariate BWAS prediction accuracies—absent overfitting—are systematically suppressed in smaller samples[5,9,21], as prediction accuracy scales with increasing sample size[1,9]. Thus, the claim that "cross-validated discovery effect-size estimates are unbiased" does not account for out-of-dataset generalizability and downward bias. In principle, if unintended overfitting and publication bias could be fully eliminated, meta-analyses of small-sample univariate BWAS would return the correct association strengths (Fig. 2f (left)). However, meta-analyses of small multivariate BWAS would always be downwardly biased (Fig. 2f (right)). If we are interested in maximizing prediction accuracy, essential for clinical implementation of BWAS[22], large samples and advancements in imaging and phenotypic measurements[1] are necessary.

Repeatedly subsampling the same dataset, as Spisak et al.[8] and we have done, overestimates reproducibility compared with testing on a truly new, diverse dataset. Just as in genomics[23], BWAS generalization failures have been highlighted[5,24]. For example, BWAS models trained on white Americans transferred poorly to African Americans and vice versa (within dataset)[24]. Historically, BWAS samples have lacked diversity, neglecting marginalized and under-represented minorities[25]. Large studies with more diverse samples and data aggregation efforts can improve BWAS generalizability and reduce scientific biases contributing to massive health inequities[26,27].

Spisak et al.[8] worry that "[r]equiring sample sizes that are larger than necessary for the discovery of new effects could stifle innovation". We appreciate the concern that rarer populations may never be investigated with BWAS. Yet, there are many non-BWAS brain–behaviour study designs (fMRI ≠ BWAS) focused on within-patient effects, repeated-sampling and signal-to-noise-ratio improvements that have proven fruitful down to $n = 1$ (ref. [28]). By contrast, the strength of multivariate BWAS lies in leveraging large cross-sectional samples to investigate population-level questions. Sample size requirements should be based on expected effect sizes and real-world impact, and not resource availability. Through large-scale collaboration and clear standards on data sharing, GWAS has reached sample sizes in the millions[29–31], pushing genomics towards new horizons. Similarly, BWAS

analyses of the future will not be limited to statistical replication of the same few strongest effects in small homogeneous populations, but also have broad scope, maximum prediction accuracy and excellent generalizability.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-05746-w.

Reporting summary

Further information on experimental design is available in the Nature Portfolio Reporting Summary linked to this Article.

Data availability

Participant-level data from all datasets (ABCD and HCP) are openly available pursuant to individual, consortium-level data access rules. The ABCD data repository grows and changes over time (https://nda.nih.gov/abcd). The ABCD data used in this report came from ABCD collection 3165 and the Annual Release 2.0 (https://doi.org/10.15154/1503209). Data were provided, in part, by the HCP, WU-Minn Consortium (principal investigators: D. Van Essen and K. Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. Some data used in the present study are available for download from the HCP (www.humanconnectome.org). Users must agree to data use terms for the HCP before being allowed access to the data and ConnectomeDB, details are provided online (https://www.humanconnectome.org/study/hcp-young-adult/data-use-terms).

Code availability

Manuscript analysis code specific to this study is available at GitHub (https://gitlab.com/DosenbachGreene/bwas_response). Code for processing ABCD data is provided at GitHub (https://github.com/DCAN-Labs/abcd-hcp-pipeline). MRI data analysis code is provided at GitHub (https://github.com/ABCD-STUDY/nda-abcd-collection-3165). FIRMM software is available online (https://firmm.readthedocs.io/en/latest/release_notes/). The ABCD Study used v.3.0.14.

1. Marek, S. et al. Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654–660 (2022).
2. Schönbrodt, F. D. & Perugini, M. At what sample size do correlations stabilize? *J. Res. Pers.* **47**, 609–612 (2013).
3. Button, K. S. et al. Confidence and precision increase with high statistical power. *Nat. Rev. Neurosci.* **14**, 585–586 (2013).
4. Varoquaux, G. Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage* **180**, 68–77 (2018).
5. Traut, N. et al. Insights from an autism imaging biomarker challenge: promises and threats to biomarker discovery. *Neuroimage* **255**, 119171 (2022).
6. Casey, B. J. et al. The Adolescent Brain Cognitive Development (ABCD) study: imaging acquisition across 21 sites. *Dev. Cogn. Neurosci.* **32**, 43–54 (2018).
7. Littlejohns, T. J. et al. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nat. Commun.* **11**, 2624 (2020).
8. Spisak, T., Bingel, U. & Wager, T. D. Multivariate BWAS can be replicable with moderate sample sizes. *Nature* https://doi.org/10.1038/s41586-023-05745-x (2023).
9. Schulz, M.-A., Bzdok, D., Haufe, S., Haynes, J.-D. & Ritter, K. Performance reserves in brain-imaging-based phenotype prediction. Preprint at https://doi.org/10.1101/2022.02.23.481601 (2022).
10. Poldrack, R. A., Huckins, G. & Varoquaux, G. Establishment of best practices for evidence for prediction: a review. *JAMA Psychiatry* **77**, 534–540 (2020).
11. Poldrack, R. A. The costs of reproducibility. *Neuron* **101**, 11–14 (2019).
12. Button, K. S. et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
13. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proc. IJCAI 95* (ed. Mellish, C. S.) 1137–1143 (Morgan Kaufman, 1995).
14. Scheinost, D. et al. Ten simple rules for predictive modeling of individual differences in neuroimaging. *Neuroimage* **193**, 35–45 (2019).
15. Sui, J., Jiang, R., Bustillo, J. & Calhoun, V. Neuroimaging-based individualized prediction of cognition and behavior for mental disorders and health: methods and promises. *Biol. Psychiatry* **88**, 818–828 (2020).
16. Woo, C.-W., Chang, L. J., Lindquist, M. A. & Wager, T. D. Building better biomarkers: brain models in translational neuroimaging. *Nat. Neurosci.* **20**, 365–377 (2017).
17. Ioannidis, J. P. A. Why most discovered true associations are inflated. *Epidemiology* **19**, 640–648 (2008).
18. Pulini, A. A., Kerr, W. T., Loo, S. K. & Lenartowicz, A. Classification accuracy of neuroimaging biomarkers in attention-deficit/hyperactivity disorder: effects of sample size and circular analysis. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **4**, 108–120 (2019).
19. Funder, D. C. & Ozer, D. J. Evaluating effect size in psychological research: sense and nonsense. *Adv. Methods Pract. Psychol. Sci.* **2**, 156–168 (2019).
20. Meyer, G. J. et al. Psychological testing and psychological assessment: a review of evidence and issues. *Am. Psychol.* **56**, 128–165 (2001).
21. He, T. et al. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *Neuroimage* **206**, 116276 (2020).
22. Leptak, C. et al. What evidence do we need for biomarker qualification? *Sci. Transl. Med.* **9**, eaal4599 (2017).
23. Weissbrod, O. et al. Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nat. Genet.* **54**, 450–458 (2022).
24. Li, J. et al. Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Sci. Adv.* **8**, eabj1812 (2022).
25. Henrich, J., Heine, S. J. & Norenzayan, A. The weirdest people in the world? *Behav. Brain Sci.* **33**, 61–83 (2010).
26. Bailey, Z. D. et al. Structural racism and health inequities in the USA: evidence and interventions. *Lancet* **389**, 1453–1463 (2017).
27. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
28. Gratton, C., Nelson, S. M. & Gordon, E. M. Brain-behavior correlations: two paths toward reliability. *Neuron* **110**, 1446–1449 (2022).
29. Levey, D. F et al. Bi-ancestral depression GWAS in the Million Veteran Program and meta-analysis in >1.2 million individuals highlight new therapeutic directions. *Nature* **24**, 954–963 (2021).
30. Muggleton, N. et al. The association between gambling and financial, social and health outcomes in big financial data. *Nat. Hum. Behav.* **5**, 319–326 (2021).
31. Yengo, L. et al. A saturated map of common genetic variants associated with human height. *Nature* **610**, 704–712 (2022).

Acknowledgements Data used in the preparation of this Article were, in part, obtained from the Adolescent Brain Cognitive Development (ABCD) Study (https://abcdstudy.org), held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children aged 9–10 years and follow them over 10 years into early adulthood. The ABCD Study is supported by the National Institutes of Health and additional federal partners under award numbers U01DA041022, U01DA041028, U01DA041048, U01DA041089, U01DA041106, U01DA041117, U01DA041120, U01DA041134, U01DA041148, U01DA041156, U01DA041174, U24DA041123, U24DA041147, U01DA041093 and U01DA041025. A full list of supporters is available online (https://abcdstudy.org/federal-partners.html). A listing of participating sites and a complete listing of the study investigators is available online (https://abcdstudy.org/scientists/workgroups/). ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in the analysis or the writing of this report. This Article reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators. Data were provided, in part, by the HCP, WU-Minn Consortium (U54 MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. This work used the storage and computational resources provided by the Masonic Institute for the Developing Brain (MIDB), the Neuroimaging Genomics Data Resource (NGDR) and the Minnesota Supercomputing Institute (MSI). The NGDR is supported by the University of Minnesota Informatics Institute through the MnDRIVE initiative in coordination with the College of Liberal Arts, Medical School and College of Education and Human Development at the University of Minnesota. This work used the storage and computational resources provided by the Daenerys Neuroimaging Community Computing Resource (NCCR). The Daenerys NCCR is supported by the McDonnell Center for Systems Neuroscience at Washington University, the Intellectual and Developmental Disabilities Research Center (IDDRC; P50 HD103525) at Washington University School of Medicine and the Institute of Clinical and Translational Sciences (ICTS; UL1 TR002345) at Washington University School of Medicine. This work was supported by NIH grants MH121518 (to S.M.), NS090978 (to B.P.K.), MH129616 (to T.O.L.), 1RF1MH120025-01A1 (to W.K.T), MH080243 (to B.L.), MH067924 (to B.L.), DA041148 (to D.A.F.), DA04112 (to D.A.F.), MH115357 (to D.A.F.), MH096773 (to D.A.F. and N.U.F.D.), MH122066 (to D.A.F. and N.U.F.D.), MH121276 (to D.A.F. and N.U.F.D.), MH124567 (to D.A.F. and N.U.F.D.), NS088590 (to N.U.F.D.), and the Andrew Mellon Predoctoral Fellowship (to B.T.-C.), the Staunton Farm Foundation (to B.L.), the Lynne and Andrew Redleaf Foundation (to D.A.F.) and the Kiwanis Neuroscience Research Foundation (to N.U.F.D.).

Author contributions Conception: B.T.-C., S.M., D.A.F. and N.U.F.D. Design: B.T.-C., S.M., R.J.C., D.A.F. and N.U.F.D. Data acquisition, analysis and interpretation: B.T.-C., S.M., R.J.C., A.V.N., B.P.K., W.K.T., T.E.N., B.T.T.Y., D.A.F. and N.U.F.D. Manuscript writing and revising: B.T.-C., S.M., R.J.C., A.V.N., B.P.K., T.O.L., W.K.T., T.E.N., B.T.T.Y., D.M.B., B.L., D.A.F. and N.U.F.D. We note that the reply author list differs from the original paper in number and in order to accurately reflect its more focused scope compared with the original work.

Competing interests D.A.F. and N.U.F.D. have a financial interest in Turing Medical and may financially benefit if the company is successful in marketing FIRMM motion monitoring software products. A.N.V., D.A.F. and N.U.F.D. may receive royalty income based on FIRMM technology developed at Washington University School of Medicine and Oregon Health and Sciences University and licensed to Turing Medical. D.A.F. and N.U.F.D. are co-founders of Turing Medical. These potential conflicts of interest have been reviewed and are managed by Washington University School of Medicine, Oregon Health and Sciences University and the University of Minnesota.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-05746-w.

Correspondence and requests for materials should be addressed to Brenden Tervo-Clemmens, Scott Marek, Damien A. Fair or Nico U. F. Dosenbach.

Reprints and permissions information is available at http://www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

nature portfolio

Corresponding author(s): Scott Marek, Brenden Tervo-Clemmens, Nico Dosenbach

Last updated by author(s): Dec 9, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection. Neuroimaging and behavioral data were from existing, open-source datasets (ABCD, UKB, HCP) whose acquisition's are presented in detail in previous work. The ABCD Study data were collected between 2016-2018. The HCP data were collected between 2010-2016.
Data analysis	MRI data analysis code can be found here: https://github.com/ABCD-STUDY/nda-abcd-collection-3165 ABCD and UKB MRI data processing code can be found here https://github.com/DCAN-Labs/abcd-hcp-pipeline Manuscript analysis code can be found here https://gitlab.com/DosenbachGreene/bwas_response FIRMM software: https://firmm.readthedocs.io/en/latest/release_notes/. ABCD uses version 3.0.14.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Participant level data from all datasets (ABCD & HCP) is openly available pursuant to individual, consortia-level data access rules. The ABCD data repository grows and changes over time. The ABCD data used in this report came from ABCD collection 3165 and the Annual Release 2.0, DOI 10.15154/1503209.

Data were provided, in part, by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. Some data used in the present study are available for download from the Human Connectome Project (www.humanconnectome.org). Users must agree to data use terms for the HCP before being allowed access to the data and ConnectomeDB, details are provided at https://www.humanconnectome.org/study/hcp-young-adult/data-use-terms.
No new data were collected for this manuscript. Across the ABCD, and HCP we downloaded data between 01/2019 - 10/2021. We did not use any specific software for downloading the data. For details on data collection in ABCD (baseline data), see Casey et al., 2018; in HCP (1200 release) see Van Essen et al., 2013).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Quantitative analyses of the magnitude and reproducibility of cross-sectional associations between neuroimaging measures and psychological/psychiatric phenotypes.
Research sample	Our main focus was to replicate work from Spisak et al in both the dataset used in their paper (HCP) and to further test generalizability of their models. Therefore, we also tested their models on the larger ABCD dataset.
Sampling strategy	All samples were recruited from the community (ABCD & HCP from the USA). Individual samples (ABCD, HCP) used unique sample size calculations and sampling strategies which are discussed in prior work with these open source datasets (Casey et al., 2018, Van Essen et al., 2013, respectively).
Data collection	All data were from existing data repositories and were downloaded between 01/2019 - 10/2021. Data used in the manuscript were from existing large consortia datasets (ABCD: see Casey et al., 2018 & Barch et al., 2018; HCP: We used data from the 1200 subjects data release (van Essen et al., 2013). Because we did not personally collect any of the data used in this manuscript, all data were from existing data repositories and researchers were therefore not blind to the source of the data.
Timing	ABCD: see Casey et al., 2018 HCP: see van Essen et al., 2013
Data exclusions	In ABCD, we used strict inclusion criteria with regard to head motion. Specifically, inclusion criteria for the current project consisted of at least 600 frames (8 minutes) of low-motion (filtered FD<0.08) resting state functional connectivity data. Our final dataset consisted of data from a total of N=3,928 youth across the discovery (N=1,964) and replication (N=1,964) sets. The final discovery and replication sets did not differ in mean FD (ΔM=0.002 , t=0.60, p=0.55) or total frames included (ΔM=6.4 , t =0.94, p=0.35). The subject lists for ARMS samples and our associated matrices will be released in the ABCD-BIDS Community Collection (ABCD collection 3165) for community use. For HCP data, we used similar data quantity inclusion, as well as an FD < 0.20 (unfiltered FD). This resulted in the inclusion of N=900 individuals (N=877 across all NIH Toolbox subscales).
Non-participation	N/A
Randomization	All three samples were observational studies and no randomization was used.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics: See above.

Recruitment: See above.

Ethics oversight: The ABCD Study obtained centralized institutional review board approval from the University of California, San Diego, and each of the 21 study sites obtained local institutional review board approval. Ethical regulations were followed during data collection and analysis. Parents or caregivers provided written informed consent, and children gave written assent.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Magnetic resonance imaging

Experimental design

Design type: resting-state fMRI, task-based fMRI; structural (cortical thickness) MRI

Design specifications: ABCD resting state: 4, 5 min runs, eyes open
HCP resting state: 4, 15 min runs, eyes open

Behavioral performance measures: Primary analyses use cognitive assessments from the NIH Toolbox and psychopathology assessment Child Behavior Checklist (see manuscript for individual subscales, total of 41) included in standard data releases and discussed in detail perviously (Barch et al., 2018)

Acquisition

Imaging type(s): Resting-state fMRI, task-fMRI, structural (cortical thickness) MRI

Field strength: 3 Tesla

Sequence & imaging parameters: Primary analyses use open-source distributed fMRI and MR data that adhere to consortia guidelines (see Casey et al., 2018 and Van Essen et al., 2013, for ABCD and HCP, respectively).

Area of acquisition: Whole brain

Diffusion MRI: ☐ Used ☒ Not used

Preprocessing

Preprocessing software: Preprocessing of ABCD was done using a suite of tools. All code can be found here: https://github.com/ABCD-STUDY/nda-abcd-collection-3165. Individual datasets (ABCD, HCP) and individual study sites (e.g., ABCD site 1 versus site 2) used unique sequence and imaging parameters which are discussed in prior work introducing these open-source datasets.

Normalization: 1) PreFreesurfer normalizes anatomical data. This normalization entails brain extraction, denoising, and then bias field correction on anatomical T1 and/or T2 weighted data. The ABCD-HCP pipeline includes two additional modifications to improve output image quality. ANTs 65 DenoiseImage models scanner noise as a Rician distribution and attempts to remove such noise from the T1 and T2 anatomical images. Additionally, ANTs N4BiasFieldCorrection attempts to smooth relative image histograms in different parts of the brain and improves bias field correction. 2) FreeSurfer 1 constructs cortical surfaces from the normalized anatomical data. This stage performs anatomical segmentation, white/grey and grey/CSF cortical surface construction, and surface registration to a standard surface template. Surfaces are refined using the T2 weighted anatomical data. Mid-thickness surfaces, which represent the average of white/grey and grey/CSF surfaces, are generated here. 3) PostFreesurfer converts prior outputs into an HCP-compatible format (i.e. CIFTIs) and transforms the volumes to a standard volume template space using ANTs nonlinear registration, and the surfaces to the standard surface

	space via spherical registration.
Normalization template	The "Vol" stage corrects for functional distortions via reverse-phase encoding spin-echo images. All resting state runs underwent intensity normalization to a whole brain mode value of 1000, within run correction for head movement, and functional data registration to the standard template (MNI). Atlas transformation was computed by registering the mean intensity image from each BOLD session to the high resolution T1 image, and then applying the anatomical registration to the BOLD image. This atlas transformation, mean field distortion correction, and resampling to 3-mm isotropic atlas space were combined into a single interpolation using FSL's 66 applywarp tool. The "Surf" stage projects the normalized functional data onto the template surfaces.
Noise and artifact removal	Additional BOLD preprocessing steps were executed to reduce spurious variance unlikely to reflect neuronal activity 46. First, a respiratory filter was used to improve FD estimates calculated in the volume ("vol") stage68. Second, temporal masks were created to flag motion-contaminated frames using the improved FD estimates 63. Frames with a filtered FD>0.3mm were flagged as motion-contaminated for nuisance regression only. After computing the temporal masks for high motion frame censoring, the data were processed with the following steps: (i) demeaning and detrending, (ii) interpolation across censored frames using least squares spectral estimation of the values at censored frames so that continuous data can be (iii) denoised via a GLM with whole brain, ventricular, and white matter signal regressors, as well as their derivatives. Denoised data were then passed through (iv) a band-pass filter (0.008 Hz<f<0.10 Hz) without re-introducing nuisance signals 69 or contaminating frames near high motion frames.
Volume censoring	Yes, ABCD data were censored at a filtered frame-wise displacement of < 0.08mm and HCP data were filtered using a non-filtered framewise displacement of <0.20mm.

Statistical modeling & inference

Model type and settings	Mass univariate and multivariate (support vector regression, canonical correlation analysis). Multiple parameterizations of each of these models were explored with the stated goal being to determine field-wide reproducibility in brain-phenotype association studies (see manuscript).
Effect(s) tested	As the primary aim of the paper was to determine the general reproducibility of brain-phenotype effects, multiple scales and combinations of effects were examined. Owing to the cross-sectional, nature of these studies, all effects are between-person associations.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☒ Both

Anatomical location(s)	Parcel-level and network-level analyses utilized the field-standard Gordon et al., 2016, Cerebral Cortex, and Seitzman et al., 2020, NeuroImage. Vertex-wise and voxel-wise data were extracted from Ciftis.
Statistic type for inference (See Eklund et al. 2016)	Multiple levels of neuroanatomical scale were used, including voxels, regions of interest, and networks.
Correction	As the primary aim of the paper was to determine the general reproducibility of brain-phenotype effects, multiple levels of significance values and correction were used, ranging from uncorrected to bonferroni (FWER) correction.

Models & analysis

n/a	Involved in the study
☐	☒ Functional and/or effective connectivity
☒	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Functional and/or effective connectivity	Pearson correlation
Multivariate modeling and predictive analysis	Two supervised regression models were used: a Ridge Regression model (α = 1.0), as proposed by Spisak et al. and a combined Principal Component Analysis (PCA) and Support Vector Regression (SVR) model, whereby half of the principal components (retaining 50% of the variance) generated from the PCA were passed as features into the SVR, as in the original work by Marek, Tervo-Clemmens et al. Both models were implemented using scikit-learn 5 in Python 3. For both HCP and ABCD datasets, both methods (ridge regression; PCA+SVR) and using three different neuroimaging feature sets (RSFC: full correlation, partial correlation; cortical thickness), the same analyses were conducted using code directly from Spisak et al. (https://gitlab.com/DosenbachGreene/bwas_response). For each behavioural phenotype and neuroimaging feature set combination, in each dataset, a complete cases sub-dataset was compiled, removing participants with missing behavioural phenotypes or neuroimaging data. For each of these complete cases (per Spisak et al.) neuroimaging feature set behavioural phenotype sub-datasets, 100 bootstraps were run for each model. Within each bootstrap, the sub-dataset was equally and randomly split into a discovery and replication set based on a given sample size. Here, sample size is defined as the size of a sole discovery/training set (identical in size to the replication set), such that given a sample size n, the total number of participants/samples of the combined discovery and replication sets is 2n. Following Spisak et al. (method and code), the discovery set was divided again into 10 cross-validation folds. However, unlike the nested cross-validation which was explored in our original manuscript and shown to not substantively change results (Marek, Tervo-Clemmens et al. 1: Supplemental Fig. S11, S12), this procedure

utilised by Spisak et al., and repeated here, did not use the additional cross-validation step for hyperparameter tuning. Rather an additional out-of-sample test was applied to the discovery dataset. TThe analyses and Figures (Fig. 1, 2) in this work use combinations of Spisak et al.'s methodological suggestions and those from our original work to replicate, expand, and clarify Spisak et al.'s Matters Arising commentary and to provide a more comprehensive perspective on out-of-sample multivariate BWAS effects. Rationale and additional details for specific analyses are provided in the relevant "Main Text" and "Figure Captions". In all cases, out-of-sample associations were evaluated as the correlation between the predicted phenotype score and the true score in the out-of-sample data. In-sample (training) associations were evaluated as the correlation between the true score and the predicted score from the model developed in the discovery set (that is, the data in the sample used to develop the model (Fig. 1).

Successful out-of-sample replication was defined as in Spisak et al.: 80% of bootstrapped iterations for a given behavioural phenotype-brain feature set ("BWAS") that were significant (via permutation test) in the first cross-validation test are significant in the second, split half test. We note this definition of replication by Spisak et al. thus does not consider all bootstrap iterations (n = 100) run when determining replication success/failure. That is, the denominator of a replication percentage is set by the number of bootstrap iterations that are significant in the first cross-validation test. Therefore, to ensure this measure of 80% replication represented a true percentage, replication here also required that more than one bootstrap iteration (out of the total 100) replicated (as defined above). Without this criteria, the impact of sampling variability and the performance of a single bootstrap iteration ensured that a small number of BWAS would appear to intermittently have replication successes followed by replication failure for the very smallest sample sizes. Reproducibility estimates following Spisak et al. guidelines were highly consistent with those from our original work.